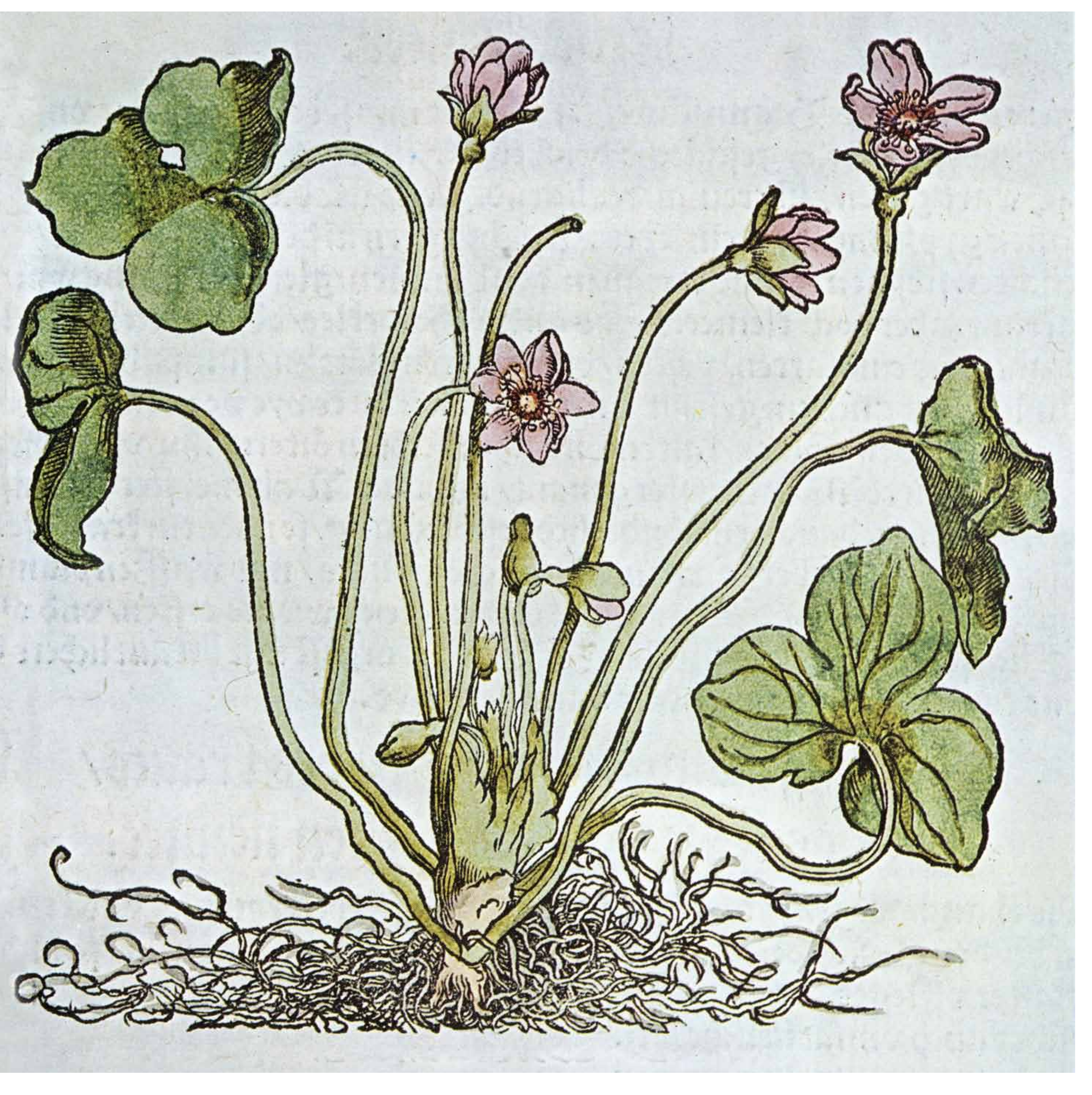
INVESTIGACION Y CIENCIA

Edición en español de SCIENTIFIC AMERICAN



LA INVESTIGACION DEL HERBARIO

Julio 1977

140 PTAS.

Los espacios en gris
corresponden a publicidad
en la edición impresa

Número 10
Julio 1977



EDITA Prensa Científica, S.A.
Calabria, 235-239
Barcelona-15
ESPAÑA

Colaboradores de este número:

Asesoramiento y traducción:

Josep Trilla: *Embalses subterráneos para el control del ciclo del agua;* José Egozcue: *Ramapithecus;* Amando García: *Semiconductores amorfos;* Juan Viñas: *Inmunología del cáncer;* M.ª Asunción Catalá: *La desaparición de las manchas solares;* Juan Antonio Seoane: *La investigación del herbario;* Luis Bou: *La paradoja de Stein en estadística;* José Enrique Myro y Alicia Menéndez: *Taller y laboratorio;* Luis Bou: *Juegos matemáticos;*

Ciencia y sociedad:

Manuel García Doncel

Libros:

Juana M.ª Golpe



LA PORTADA

La ilustración de la portada reproduce un grabado sacado de un herbolario alemán, de rara difusión, *Contrafayt Kreüterbuch* ("libro de plantas ilustrado"). Se imprimió en Estrasburgo entre 1532 y 1537, y se conserva hoy en la biblioteca del New York Botanical Garden. Otho Brunfels, que era médico y botánico, escribió el texto; las ilustraciones procedían de grabados de madera que se basaron en las acuarelas de Hans von Weiditz. La portada representa el *Leberkraut* (hierba para el hígado), conocida en la actualidad por *Hepatica triloba* o hepática. Los herbolarios eran compendios del conocimiento contemporáneo sobre materia vegetal haciéndose especial hincapié en sus aplicaciones terapéuticas. De la hepática se decía, por ejemplo, que "curaba la bilis amarilla". Hasta tiempos muy recientes, la medicina iba de la mano de la botánica: incluso en nuestros días la mayoría de las drogas proceden de substancias vegetales. De ahí que, cara al futuro, un recurso valioso a explotar será el de los herbarios existentes en todas partes, en donde se conservan, con los centenares de miles de especies, notas etnobotánicas que hablan de su uso terapéutico en las poblaciones indígenas (véase "La investigación del herbario", en este mismo número).

Suscripciones:

Prensa Científica, S.A.
Departamento de suscripciones
Calabria, 235-239
Barcelona-15 (España)

Distribución para España:

Distribuciones de Enlace, S.A.
Ausias March, 49
Barcelona-10 (España)

Distribución para los restantes países:

Editorial Labor, S.A.
Calabria, 235-239
Barcelona-15 (España)

Condiciones de suscripción:

España:
Un año (12 números): 1.540 pesetas

Extranjero:
Un año (12 números); 30 U.S. $

Ejemplar atrasado ordinario:
160 pesetas

Ejemplar atrasado extraordinario:
225 pesetas.

PROCEDENCIA DE LAS ILUSTRACIONES

Fotografía de la portada por Fritz Goro

Página	Fuente	Página	Fuente
4-10	Ilil Arbel	49-50	Patricia J. Wynne
13-18	Tom Prentiss	51	William D. Hardy, Jr., Sloan-Kettering Institute for Cancer Research
19	Elwyn L. Simons, Yale University	52-53	Gabor Kiss
23	Fritz Goro	54	Lloyd J. Old, Sloan-Kettering Institute for Cancer Research
24-34	Dan Todd	59	New York Public Library
39	Dokumente der Naturwissenschaft	60-66	Allen Beechel
43	Andrejs Liepins, Sloan-Kettering Institute for Cancer Research	71-73	Fritz Goro
44-47	Patricia J. Wynne	76	The Bettmann Archive
48	Etienne de Harven, Sloan-Kettering Institute for Cancer Research (*arriba*), Patricia J. Wynne (*abajo*)	81-85	Federico Mayor
		87-88	Scientific American
		90-92	Federico Mayor
		95-101	Gabor Kiss
		105-108	Ilil Arbel

Dep. legal: B. 38.999-76
Fotocomposición: Fort, S.A.
Rosellón, 33 - Barcelona-15
Cayfosa. Santa Perpetua de Moguda. Barcelona.
Printed in Spain - Impreso en España

Los autores

ROBERT P. AMBROGGI ("Embalses subterráneos para el control del ciclo del agua"). Desde 1961 ha sido consejero de la Organización para la Alimentación y la Agricultura de las Naciones Unidas (FAO). Trabaja además como consultor en el Programa de Desarrollo y en el Programa del Medio Ambiente de las Naciones Unidas, así como en el Banco Mundial. Se graduó en 1939 por la Universidad de Nancy, doctorándose por la Sorbona de París. Antes de incorporarse a la Organización para la Alimentación y la Agricultura dirigió el Departamento de Recursos Hidráulicos de Marruecos, en donde empezó a trabajar en 1942 en el campo de la hidrogeología del agua subterránea. El principal interés de Ambroggi se centra en la adecuada gestión de los escasos recursos de agua de buena calidad. Participa como miembro activo del Plan Azul del Mediterráneo, acción regional de los 18 países costeños para asesorar a sus gobiernos en el desarrollo y aprovechamiento nacional de los recursos naturales.

ELWYN L. SIMONS ("Ramapithecus") enseña geología en la Universidad de Yale y dirige el departamento de paleontología de vertebrados del Yale Peabody Museum. Se le concedió, el año pasado, el premio Alexander von Humboldt de la República Federal Alemana. Coincidiendo con su año sabático ha proseguido sus estudios sobre *Ramapithecus* en Europa, Africa y Oriente Próximo. Ha dirigido o ha participado en más de 30 expediciones en busca de fósiles por los tres continentes; en su haber cuenta además con dos libros y más de 100 comunicaciones de investigación. De sus hallazgos, el más sobresaliente es el del *Aegyptopithecus* y otros antropoides del Oligoceno norteafricano, así como el descubrimiento de la primera muestra de *Gigantopithecus* al norte de la India.

DAVID ADLER ("Semiconductores amorfos") es profesor de ingeniería eléctrica en el Instituto de Tecnología de Massachusetts. Tras asistir a la Bronx High School of Science y al Rensselaer Polytechnic Institute, se graduó en la Universidad de Harvard en 1964. Antes de incorporarse al M.I.T., permaneció un año en el Atomic Energy Research Establishment de Harwell, en Inglaterra. Es miembro del Comité ejecutivo de la Division of Solid State Physics de la American Physical Society y forma parte del consejo editorial de *Semiconductors and Insulators.* Ha trabajado también en el Basic Research Committee que asesora a la U.S. Army Research Office en el National Research Council, presidió el Solar Photovoltaic Panel del M.I.T. Solar Energy Workshop y el Harvard Graduate Student Council.

LLOYD J. OLD ("Inmunología del cáncer") es vicepresidente del Sloan-Kettering Institute para la investigación del cáncer. Nació en California y se graduó en la Universidad de Berkeley. En 1958 se doctoró en medicina. Año en que pasó al Sloan-Kettering Institute. En 1973 fue nombrado vicepresidente y director adjunto de investigación del Memorial Sloan-Kettering Cancer Center and Memorial Hospital for Cancer and Allied Diseases. Ha enseñado biología en la Sloan-Kettering Division de la Facultad de Medicina de la Universidad de Cornell. Old es uno de los fundadores de la inmunología tumoral científica. Fue de los primeros en presentar pruebas de la habilidad que tienen algunos derivados microbianos para aumentar de modo general la capacidad de respuesta inmunológica y de incrementar, por tanto, la resistencia frente a las células cancerosas.

JOHN A. EDDY ("La desaparición de las manchas solares") es astrónomo especialista en física solar del Center for Astrophysics del Harvard College Observatory y del Smithsonian Astrophysical Observatory. Se graduó en la U.S. Naval Academy en 1953 y se doctoró por la Universidad de Colorado en 1962. Al año siguiente estuvo, como investigador en radioastronomía, en el National Bureau of Standards. En 1963 entró a formar parte del personal investigador del High Altitude Observatory en Colorado; al mismo tiempo impartió cursos de astronomía y de historia de la astronomía en la Universidad de Colorado. Entre 1967 y 1970 fue asesor del College Science Improvement Program. Además de trabajar en astronomía solar, Eddy también ha investigado en astronomía infrarroja, historia de la astronomía y astronomía arqueológica.

SIRI VON REIS ALTSCHUL ("La investigación del herbario") es directora honoraria de etnobotánica en el New York Botanical Garden e investigadora honoraria en etnofarmacología en el Botanical Museum de la Universidad de Harvard. Tras licenciarse en botánica por la Universidad de Michigan en 1953, se trasladó a Harvard para doctorarse en 1961. Entre 1962 y 1972 trabajó como profesora de investigación en el Harvard Botanical Museum.

FEDERICO MAYOR ("Bioquímica de la subnormalidad") es catedrático y director del departamento de Bioquímica de la Universidad Autónoma de Madrid. Dirige, asimismo, el Centro de Biología Molecular y preside la Sociedad Española de Bioquímica. En su extensa producción científica se distinguen tres periodos: de 1956 a 1963, en que trabaja e investiga sobre la glutamato descarboxilasa; de 1963 a 1972, en que se centra sobre el diagnóstico precoz de enzimopatías, y de 1972 hasta nuestros días en que ha insistido en el análisis bioquímico de las causas del retraso mental. Pertenece a numerosas instituciones europeas y norteamericanas.

BRADLEY EFRON y CARL MORRIS ("La paradoja de Stein en estadística") son, respectivamente, jefe del departamento de estadística de la Universidad de Stanford, y estadístico de plantilla de la Rand Corporation de Santa Mónica. Ambos recibieron su primera formación universitaria en el Instituto de Tecnología de California, continuando estudios en el departamento de estadística de la Universidad de Stanford. Efron se doctoró en 1964; Morris, en 1966, bajo la dirección de Charles Stein (el descubridor de la paradoja de su nombre). Efron es también profesor de medicina preventiva en la Facultad de Medicina de la Universidad de Stanford, trabajando en el estudio de datos médicos. Además, durante tres años, dirigió la sección de teoría y métodos del *Journal of the American Statistical Association.* Morris ha permanecido en la Rand Corporation desde 1967, donde trabaja en el diseño estadístico del Health Insurance Study, patrocinado por el Department of Health, Education and Welfare, al objeto de determinar los efectos de diferentes programas de seguridad sanitaria en la demanda de servicios médicos y en el estado general de salubridad. Antes de incorporarse a la Rand, enseñó estadística en la Universidad de California en Santa Cruz, en la de Stanford y en el Rand Graduate Institute.

Embalses subterráneos para el control del ciclo del agua

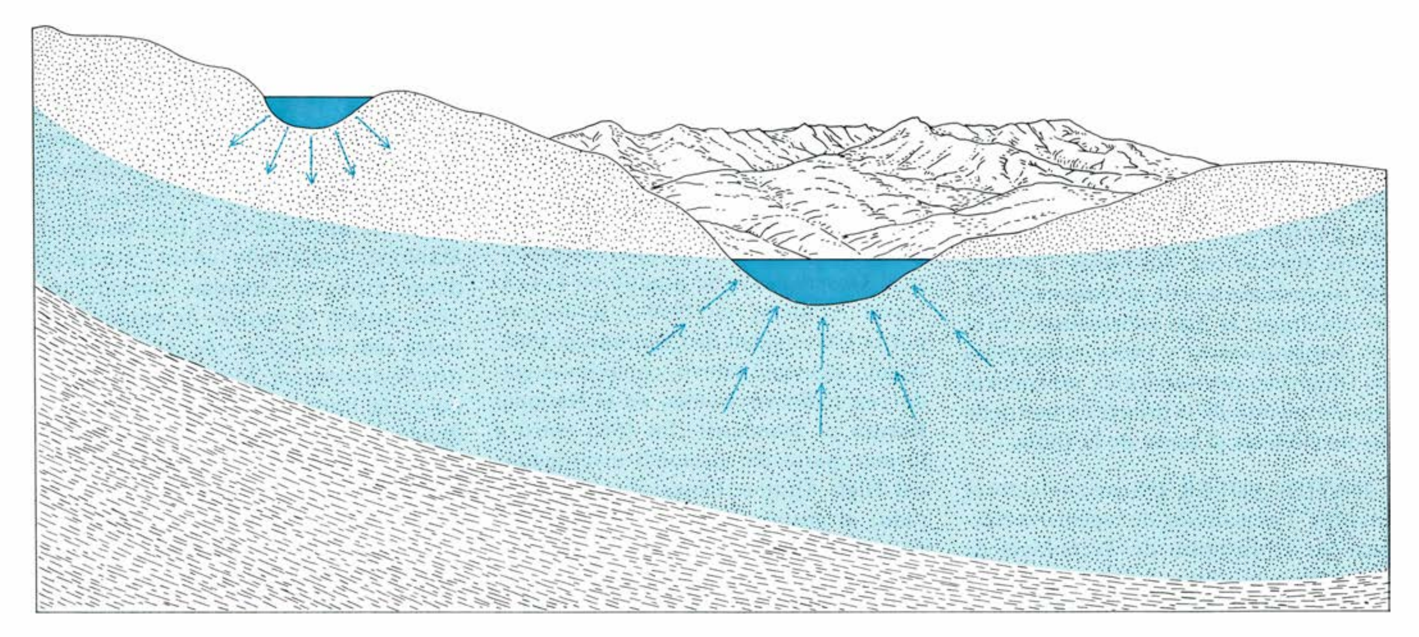
Las dos terceras partes del agua dulce existente en la tierra se encuentran retenidas en embalses subterráneos. Estos podrían explotarse más y mejor y rellenarse cuando abundara el agua

Robert P. Ambroggi

Se tiende a pensar que el ciclo global del agua procede de la manera siguiente. La energía solar evapora el agua de los océanos; los procesos meteorológicos la transportan a través de la atmósfera, desde la cual se precipita sobre la tierra en forma de lluvia o nieve, para volver finalmente a los océanos a través de ríos y arroyos.

Esta visión sumaria pasa por alto las enormes cantidades de agua que residen durante diversos periodos de tiempo (desde semanas hasta milenios) en los embalses subterráneos. Además, muy pocos de los responsables de la planificación de la distribución de los recursos del agua han valorado adecuadamente la justa importancia de los embalses subterráneos a la hora de hacer frente a la escasez de agua, escasez que cada vez acosa más a la agricultura y a otras actividades humanas en varias partes del mundo; e, igualmente, su destacado papel a la hora de incrementar grandemente la cantidad de agua que podría ponerse al servicio del esfuerzo humano. En este artículo sostengo que el género humano debería explotar racionalmente los embalses subterráneos –mucho más de lo que en algún grado ya hace– para satisfacer las necesidades de la agricultura, la industria y las actividades sociales, ya que, tarde o temprano, la naturaleza, ayudada a veces por la mano del hombre, volverá a llenar los embalses. Se obtendrá así un control muchísimo mayor del ciclo del agua, evitándose las crisis por causa de ésta a que habrían de enfrentarse prácticamente todas las naciones en las próximas décadas, de continuar las actuales prácticas derrochadoras en la administración del agua.

La escasez general de cosechas de 1972 fue un avance de las dificultades que nos esperan si el agua no se administra con mayor eficiencia. Las carestías se debieron a un déficit de unos 400 kilómetros cúbicos (400.000 millones de



LOS EMBALSES SUBTERRANEOS naturales se llenan cuando el agua se infiltra a través de suelo o a través de formaciones rocosas permeables por debajo de un río o arroyo. Puede permanecer allí a lo largo de periodos que oscilan desde semanas hasta milenios, antes de encontrar su camino de vuelta al ciclo global de evaporación, precipitación y escorrentía. El agua subterránea constituye también una parte significativa del flujo de la mayoría de los arroyos. Entre otros procesos naturales que contribuyen a la realización de los embalses subterráneos están la lluvia y la nieve que caen sobre el terreno, y la infiltración desde lagos y pantanos. Los embalses agotados por el uso intensivo pueden rellenarse artificialmente mediante la infiltración desde canales y lagos construidos por el hombre y la inyección de agua en el terreno.

metros cúbicos) en la parte continental del ciclo del agua. A escala del flujo global del agua, el déficit fue pequeño, pero tuvo grave repercusión, porque la mayoría de los países usan el agua de manera tal que dejan un margen pequeño para abordar las consecuencias de la sequía. De mantenerse las prácticas actuales, la producción de alimentos en el mundo se irá acercando peligrosamente al límite de su autocapacidad. La prevención de la crisis de alimentos requerirá, en el futuro, una manipulación del agua a una escala mucho mayor que la que las naciones han practicado hasta ahora. Se ve venir la necesidad de un notable aumento de la irrigación y el trasvase de agua entre cuencas hidrográficas y aun entre continentes.

El personal que entiende en la gestión del agua distingue entre recursos hídricos renovables y recursos no renovables. En el ciclo convencional del agua que he descrito se mueven los recursos renovables, que ascienden a unos 500.000 kilómetros cúbicos por año. Los recursos no renovables tienen un volumen de 1500 millones de kilómetros cúbicos. Están almacenados en tres embalses: los océanos (97,3 por ciento de los recursos), los continentes (2,7 por ciento) y la atmósfera (0,0001 por ciento).

Estas distribuciones pasan por alto el hecho de que todos los recursos en agua de un continente provienen de un ciclo. La única diferencia entre recursos renovables y no renovables es la rapidez con que el agua se recicla. Los llamados recursos renovables se mueven bastante rápidamente, mientras que los recursos no renovables se mueven con extrema lentitud.

El periodo de tiempo que tarda una cantidad particular de agua en su ciclo se llama su tiempo de residencia. El tiempo de residencia del agua en el embalse atmosférico puede ser de 10 días. En la profundidad de los océanos, el tiempo de residencia de una cantidad de agua puede durar más de 1000 años. El tiempo de residencia de los recursos de agua continental oscila desde unas cuantas semanas hasta muchos milenios, dependiendo del tipo de embalse: cauce de un río, lago, pantano, glaciar, hielo polar y agua subterránea. Los embalses de agua subterráneos muestran la mayor diversidad de tiempo de residencia: unos pocos días o semanas en acuíferos cársticos (calizas); unas cuantas semanas o meses en cauces gravosos; algunos meses o años en depósitos aluviales; unos cuantos años o décadas en otros acuíferos no confinados, y varias décadas, siglos o milenios en acuíferos confinados.

Examinemos más de cerca el movimiento del agua en el ciclo hidrológico convencional. La evaporación anual desde los océanos y continentes es de unos 500.000 kilómetros cúbicos por año: 430.000 desde los océanos y 70.000 desde los continentes. El otro miembro de la ecuación es la precipitación, que, en promedio, alcanza 110.000 kilómetros cúbicos por año sobre los continentes y 390.000 kilómetros cúbicos sobre los océanos. Expresado de otra forma, los océanos suministran 430.000 kilómetros cúbicos por evaporación y reciben 390.000 por precipitación; por su parte, los continentes pierden 70.000 por evaporación y reciben 110.000 por precipitación. Por lo tanto, cada año se trasladan 40.000 kilómetros cúbicos de agua dulce desde los océanos a los continentes. De acuerdo con el principio de la conservación del agua dentro del ciclo hidrológico, cada año vuelven a los océanos 40.000 kilómetros cúbicos en forma de escorrentía superficial y subterránea.

Por consiguiente, se puede suponer que serían aprovechables para la humanidad 40.000 kilómetros cúbicos por año, si los recursos hídricos fueran utilizados con la máxima eficiencia. De hecho, gran parte de esta agua resulta inaccesible: se encuentra en avenidas, agua retenida en el suelo, escorrentías en lugares inhabitados y otros casos parecidos. La humanidad puede aspirar a utilizar, como cota máxima, unos 14.000 kilómetros cúbicos, que es el flujo de base, o estable, en ríos y arroyos. El flujo de base está regulado principalmente por las descargas desde embalses subterráneos (11.000 kilómetros cúbicos), desde embalses construidos por el hombre (1840) y desde lagos (260).

Tampoco el total de los 14.000 kilómetros cúbicos está completamente al alcance de la mano, a menos que los transportes del agua se hagan a una escala enorme, ya que gran parte del total se halla en zonas inhabitables. Hacia 1972, la cantidad de agua bajo control humano era aproximadamente de 3000 kilómetros cúbicos por año. Teniendo en cuenta que unos 5000 kilómetros cúbicos de la escorrentía fluvial estable se produce en regiones inhospitalarias, subsisten unos 6000 kilómetros cúbicos que, bien administrados, el hombre podría emplear para su provecho en el futuro.

La especie humana sólo habita, normalmente, 60 de los 150 millones de kilómetros cuadrados de los continentes; se halla realmente excluida de lugares tales como los desiertos, las regiones polares, las altas montañas y las selvas. De las zonas habitables, unos 3200 millones de hectáreas de terreno son potencialmente útiles para la agricultura; en 1970 se cultivaban unos 1400 millones de hectáreas. No es razonable suponer que la cantidad de terreno adicional que se podría someter a cultivo durante el próximo siglo llegue a los 1000 millones de hectáreas. Para satisfacer la necesidad creciente de alimento constituiría una estrategia mejor usar el terreno más intensivamente, para lo cual el riego desempeñaría un papel principal. La extensión de tierras regadas en 1970 abarcaba 180 millones de hectáreas. El riego consumía unos 2400 kilómetros cúbicos de agua por año. La industria consumía 500 y otras actividades humanas 200; el consumo de agua de toda la humanidad era, pues, de 3100 kilómetros cúbicos.

La crisis alimentaria de 1972 surgió de un déficit de 35 millones de toneladas métricas de cereales (trigo, arroz y granos comunes). Como consecuencia adicional, las existencias de trigo de los países exportadores bajaron de 50 millones de toneladas métricas en 1971 a 30 millones en 1974. Otro año o dos más de déficit en la precipitación crearía una crisis alimentaria global que acarrearía dramáticas consecuencias internacionales. Así, parece que la situación de la alimentación en el mundo ha alcanzado un punto en el cual los efectos de la deficiencia de agua son mucho más graves y sentidos que en el pasado.

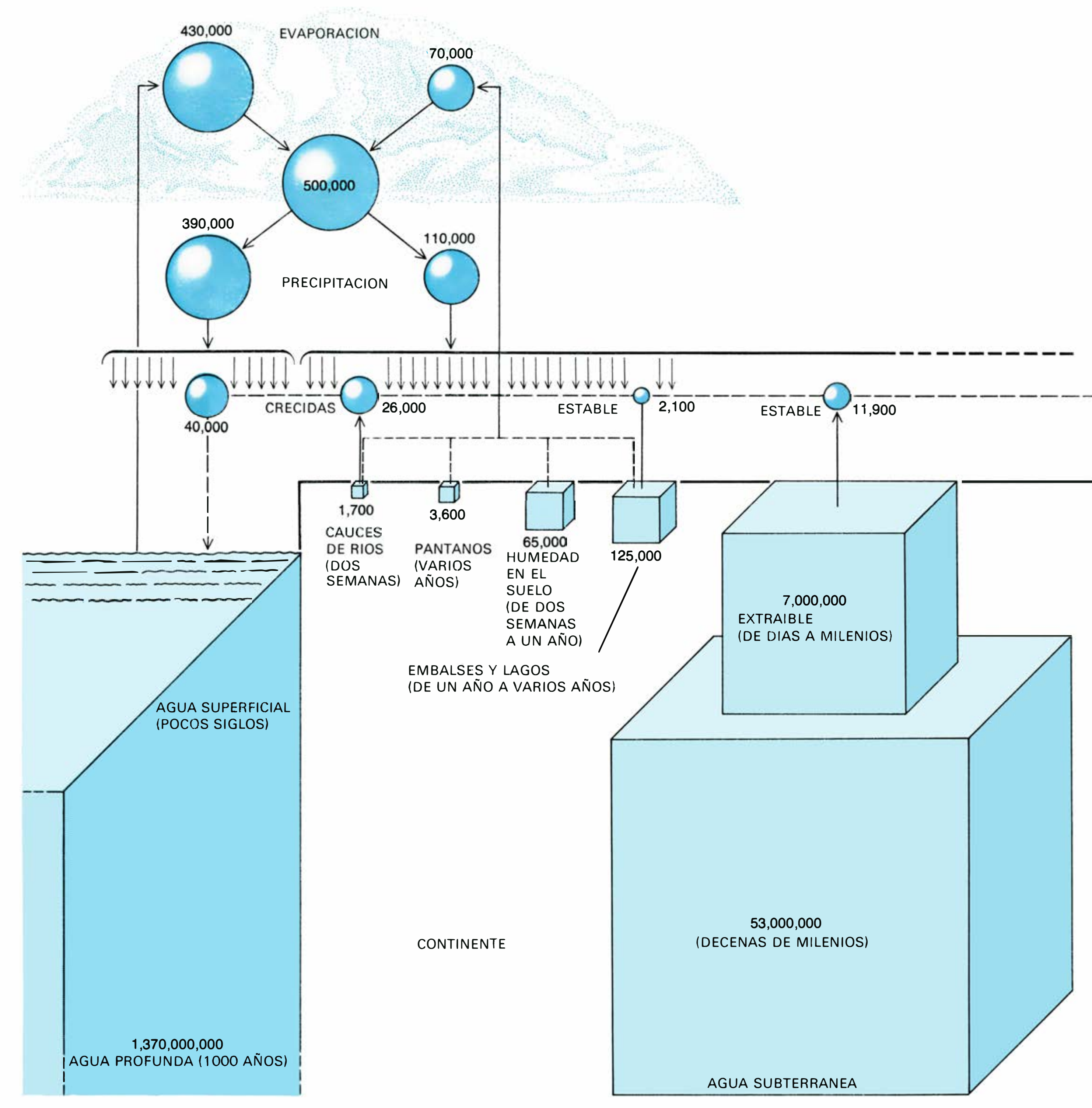
Otro problema inquietante en la producción de alimentos es la posibilidad de que se hallen cambiando las pautas climáticas del mundo. Un ejemplo lo ofrece la gran cobertera de vientos de elevada altitud que giran alrededor del polo de oeste a este, en el vórtice circumpolar nórdico. En los últimos años, el límite más bajo del vórtice nórdico se ha registrado más al sur de lo que estuvo algunos años atrás. A consecuencia de ello, las zonas de alta presión se desplazaron también más al sur, bloqueando así las lluvias monzónicas de las regiones donde son vitales para la supervivencia de cientos de millones de personas. Las lluvias cayeron, por tanto, en los océanos o sobre zonas continentales cuya pluviosidad era ya más que suficiente.

Todos estos hechos suscitan la cuestión de si el hombre es tecnológicamente capaz de manipular el ciclo del agua, hasta el punto de evitar las calamidades que puedan surgir por la escasez de lluvia prolongada o por cambios de clima. Se han llevado a cabo, o se han propuesto, un buen número de esfuerzos

técnicos para abordar tales problemas; por ejemplo, producción de nubes, transporte a larga distancia de icebergs, desalinización del agua del mar y almacenamiento del agua fluvial mediante sistemas de presas y embalses.

El esfuerzo para inducir la precipitación "sembrando" nubes con yoduro de plata, dióxido de carbono helado y otras substancias, no dispone todavía de una firme base científica. Sigue siendo un empeño un tanto caro y aleatorio. De los muchos experimentos en este campo sólo unos cuantos han demostrado que la producción de nubes puede aumentar la precipitación, dentro de un limitado margen de condiciones favorables.

Se han propuesto remolcar icebergs polares, principalmente desde la Antértida, hasta las costas continentales áridas. Así, por ejemplo, se podría remolcar un gran iceberg (que contenga 6 kilómetros cúbicos de agua dulce) desde la Antártida al desierto de Atacama en Chile en unos siete meses, con una pérdida de agua del 30 por ciento. El iceberg sería amarrado en su lugar de destino y después fundido o cuarteado para varios fines. Aunque no se ha hecho realmente nada en esta línea, valdría la pena considerar un proyecto piloto para países en vías de desarrollo, como Chile



LOS RECURSOS HIDRICOS del mundo están representados en este diagrama que muestra el ciclo anual *(arriba)*, el agua almacenada en los océanos *(abajo, a la izquierda)* y en los continentes *(derecha)*. Los números expresan kilómetros cúbicos de agua. También se indica el periodo probable de tiempo de permanencia del agua en determinada zona de almacenamiento. 430.000 kilómetros cúbicos de agua se evaporan de los océanos y 70.000 de los continentes; la precipitación es de 390.000 kilómetros cúbicos por año sobre los océanos y 110.000 sobre los continentes. Se produce así una transferencia de 40.000 kilómetros cúbicos de agua dulce por año desde los océanos a los continentes, que

y Perú. La descarga del Antártico asciende a unos 2000 kilómetros cúbicos de icebergs por año, un recurso hídrico considerable, por tanto.

La desalinización del agua de mar ha atraído mucho la atención, particularmente porque durante la pasada década el coste se ha reducido desde aproximadamente 1,50 dólares por 1000 galones (3785 litros) hasta 50 centavos (es decir, 13 centavos por metro cúbico). La tendencia actual es construir plantas nucleares con el doble fin de generar electricidad y desalinizar grandes cantidades de agua. Pero este planteamiento sólo parece factible para el abastecimiento doméstico de agua en regiones costeras de países ricos, dada la inversión de capital requerida y el elevado coste del funcionamiento de las plantas.

Hasta ahora, el método más eficaz empleado por la humanidad para manipular el ciclo del agua ha sido la construcción de presas y embalses que regulen el ciclo del flujo de los ríos. Hay unos 1840 kilómetros cúbicos de agua dulce almacenados, que se distribuyen así: 560 en Asia, 490 en Norteamérica, 400 en Africa, 200 en Europa, 160 en Sudamérica y 30 en Australia. Las perspectivas para una gran expansión de esta tecnología parecen ser, sin embargo, limitadas, especialmente en los países en desarrollo, ya que el coste de regular un kilómetro cúbico por año por el método de presas de embalse asciende ahora a unos 100 millones de dólares. Por otra parte, el tiempo de residencia del agua almacenada raramente excede de un año; de ahí que el sistema ofrezca una protección inadecuada contra un déficit de lluvia que perdure más de un año.

A pesar de eso, al construir sistemas de presas de embalse, la humanidad ha puesto en movimiento, inconscientemente, la regulación a largo plazo del ciclo del agua a través de los embalses de agua subterránea. Por ejemplo, el vasto sistema de riego de la cuenca del río Indo en el Pakistán, que se inició en 1860, originó una densa red de canales de riego. Una gran proporción del agua superficial de los ríos desviada a estos canales se infiltró hacia los embalses subterráneos. El resultado fue que la superficie piezométrica del agua ascendió constantemente durante más de un siglo, creando un gran almacenamiento subterráneo. Empresas de esta clase permiten pensar en un control del ciclo del agua a largo plazo mediante la recarga artificial de embalses subterráneos. ¿Es ésta una estrategia solvente para el futuro?

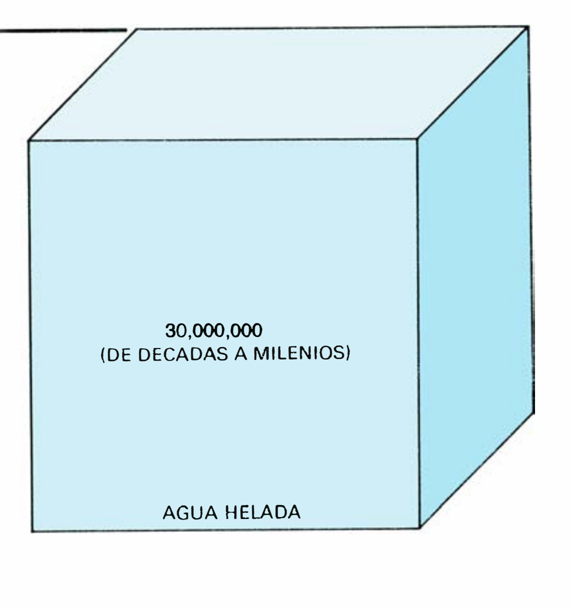
La naturaleza recicla anualmente unos 12.000 kilómetros cúbicos de agua a través de los embalses subterráneos para constituir el flujo estable o bajo de los ríos; esto representa el 30 por ciento de la escorrentía total. La humanidad está regulando, mientras tanto, menos de 2000 kilómetros cúbicos por sistemas de presas de embalse. Por otra parte, sólo 1200 kilómetros cúbicos de agua de los 3100 que la humanidad usa anualmente proceden de embalses subterráneos. Durante más de un siglo los gobiernos han confiado en la solución consistente en presas de embalse y se han despreocupado olímpicamente de los recursos subterráneos disponibles.

Los embalses subterráneos tienen varias funciones, entre las cuales se cuenta el suministro, almacenamiento, mezcla y conducción del agua. Hasta ahora el ser humano ha confiado principalmente en la función de suministro, esto es, en la perforación de pozos que explotan las existencias subterráneas. Sólo en California, y más recientemente en Israel, se ha puesto mucha atención en la función de almacenamiento. California desvía gran cantidad de agua desde el norte, donde se encuentra en abundancia, hacia el sur, en donde escasea. El agua subterránea se ha agotado extensamente en el sur a causa del bombeo, y gran parte del agua trasladada desde el norte sirve para rellenar esos embalses subterráneos. Los embalses agotados proporcionan, por consiguiente, un mecanismo de almacenamiento.

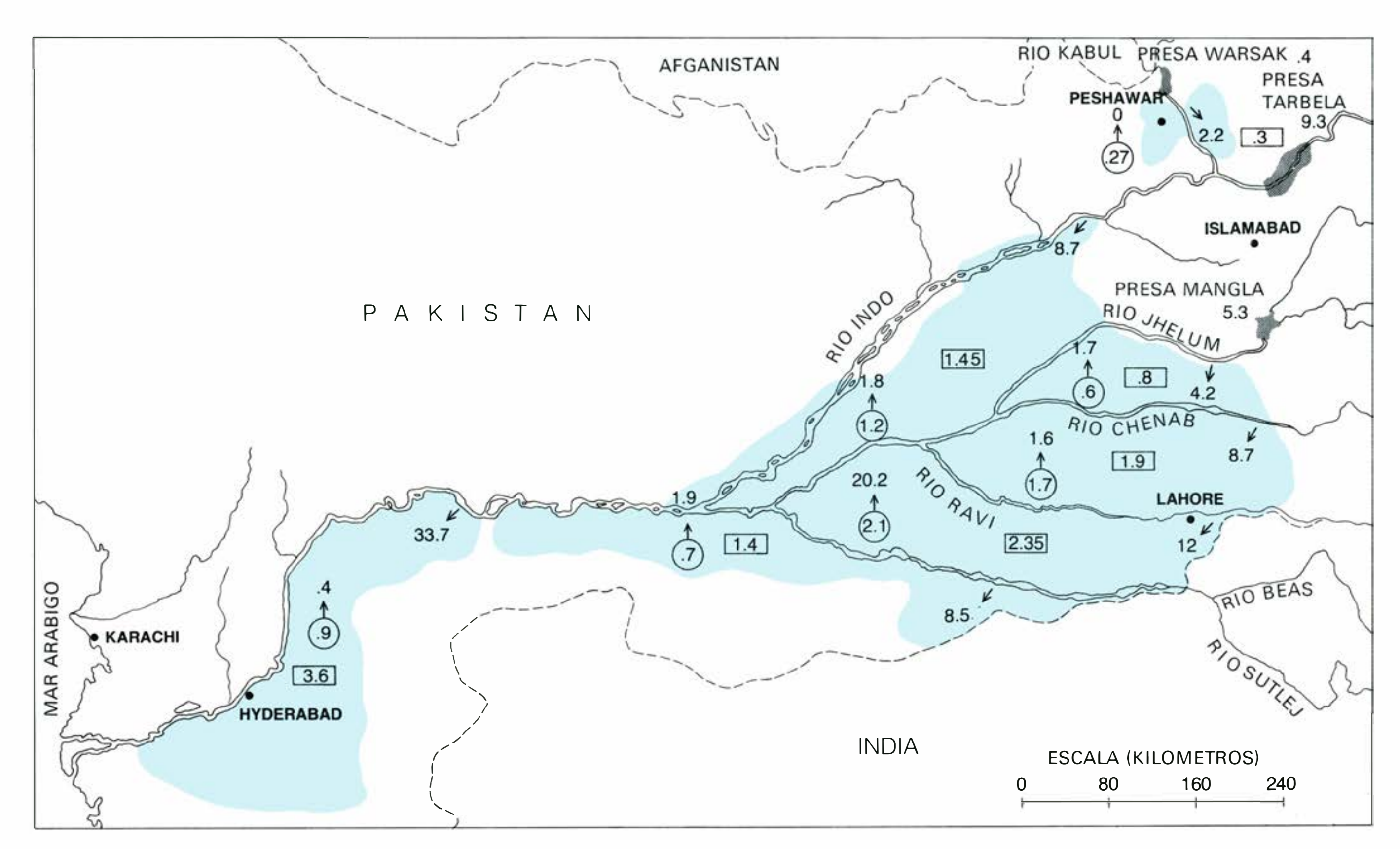
En Israel, de manera análoga, el sistema denominado Transporte Hídrico Nacional traslada 300 millones de metros cúbicos de agua por año del norte al sur. De ellos, unos 200 millones de metros cúbicos se almacenan en invierno en los principales embalses de agua subterránea de areniscas y calizas. Los objetivos son cubrir la demanda máxima de agua en verano y la mayor demanda en años de precipitación escasa, así como también mejorar la calidad del agua mediante mezcla.

Obviamente, un embalse de agua subterránea que esté lleno no puede emplearse para almacenamiento adicional a largo plazo. Esta era la situación de casi todos los embalses subterráneos del mundo hasta el siglo XIX, en que se difundió la tecnología de perforación y de bombeo. No obstante, solamente unos cuantos embalses de agua subterránea siguen estando lo bastante agotados como para proporcionar espacio para atrapar agua del ciclo hidrológico. Parte del problema reside en que los hidrólogos se muestran generalmente conservadores e insisten en que la explotación de los embalses de agua subterránea se mantenga menor que la "producción de seguridad", que se define como la cantidad de producción que puede esperarse sea reemplazada por recarga natural durante el año.

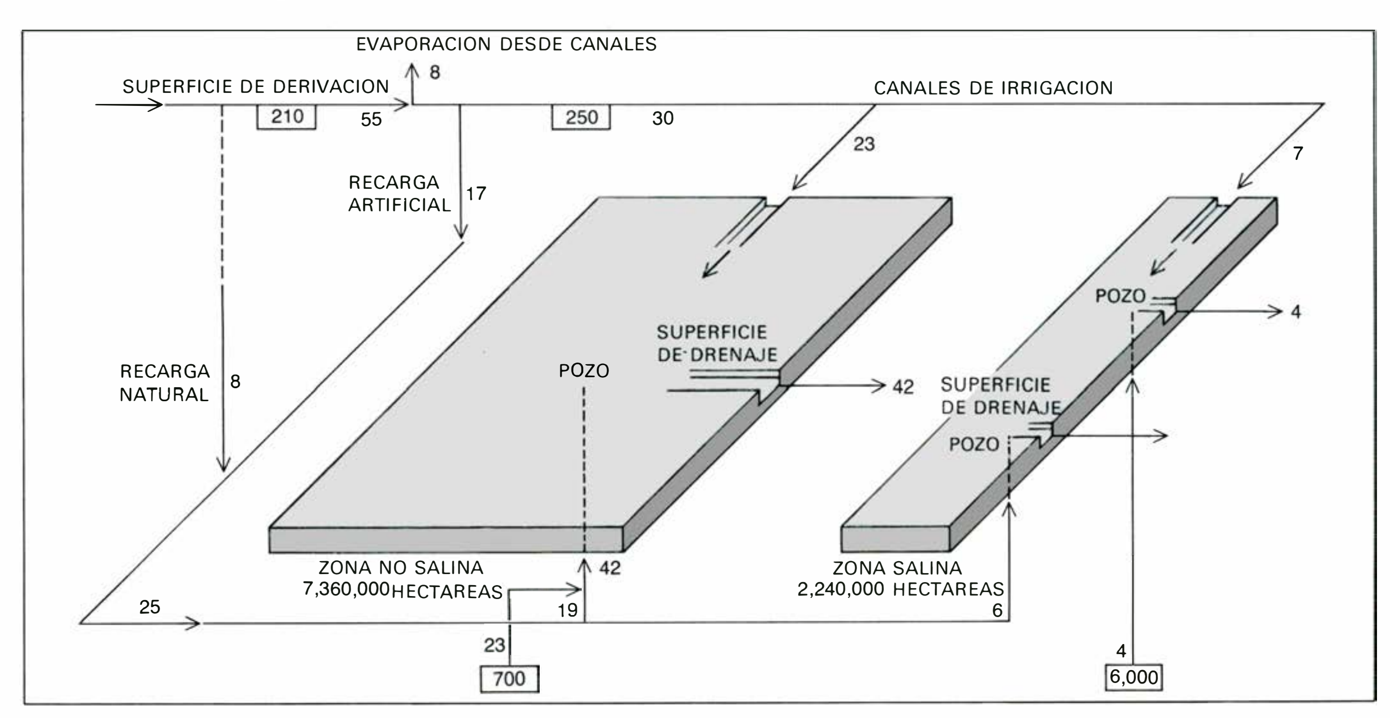
Todo embalse administrado de esta manera debe considerarse como un embalse saturado. No hay espacio para recarga deliberada de almacenamiento a largo plazo, que es la mejor manera de regular el ciclo del agua. Por consiguiente, la primera condición para el almacenamiento a largo plazo es au-



la escorrentía compensa en el ciclo anual. Todas las zonas de almacenamiento continental contribuyen a la escorrentía, como muestran los números próximos a los círculos. Se omite la escorrentía desde los hielos polares.



EL SISTEMA DE RIEGO de la cuenca de río Indo en Pakistán representa una enorme aportación humana al embalse de agua subterránea a través de un sistema de 60.000 kilómetros de canales dispuestos en las regiones señaladas en color. Los números en rectángulos de las regiones canalizadas representan, en millones de hectáreas, la extensión de tierra afectada por el sistema de canales; los números en círculos muestran, en millones de hectáreas, el área bajo la cual hay agua subterránea de calidad útil. Las cifras restantes representan kilómetros cúbicos de agua; quedan explicadas en la leyenda de la izquierda. La construcción del sistema empezó en 1860. Gradualmente, la infiltración de agua desde los canales hacia el terreno elevó la superficie piezométrica de modo que empezó a interferir con la agricultura. Ahora, el sistema se administra de forma que se bombea deliberadamente parte del agua subterránea, para riego.



GESTION DEL AGUA en el valle del Indo; esquema a partir de una idea de H. A. Thomas, Jr., y Robert P. Burden, de la universidad de Harvard. Las cifras asociadas con flechas indican flujo de agua en kilómetros cúbicos por año. Las cifras en rectángulos representan (en partes por millón) la concentración del total de sólidos disueltos en el agua, es decir, su salinidad. El total de los 9,6 millones de hectáreas de tierra regada está dividido en zona salina y zona no salina, porque los acuíferos se administran de modo diferente en cada caso. La capa superior del embalse de agua subterránea se bombea en exceso respecto de la recarga anual, haciendo descender así el nivel elevado de la superficie piezométrica, que había subido durante la pasada centuria. Ello produce una disminución de los problemas de anegamiento y salinidad que habían obstaculizado la agricultura. Además, el uso del agua subterránea en el riego ha permitido aumentar notablemente la zona cultivada.

mentar deliberadamente la explotación de un acuífero, más allá de la producción de seguridad. Si es posible, la explotación debería también hacer descender el nivel del embalse hasta que quede por debajo de sus salidas naturales. En resumen, la política sería sobreexplotar el embalse.

Tal política implica una atención al problema de recarga del embalse, que puede hacerse ya sea natural o artificialmente. La recarga natural de un embalse de agua subterránea resulta de la percolación descendente del agua de los ríos y también del agua que cae en forma de lluvia o de nieve sobre suelos permeables. La experiencia recogida durante las tres últimas décadas ha demostrado que un solo año de lluvia extraordinariamente abundante, cuya frecuencia es de, al menos, una vez cada 15 años, rellenará un acuífero cuyo nivel hubiera descendido de 10 a 12 metros en los años anteriores. Por ejemplo, las avenidas de 1969 en Túnez rellenaron todos los embalses que, hasta entonces, se habían descrito como superexplotados. Similarmente, un acuífero poco profundo en el valle Souss de Marruecos había soportado una floreciente producción de cítricos desde 1946; pero hacia 1957 fue motivo de gran inquietud porque la superficie piezométrica (water table) había descendido un metro por año. Las lluvias de 1957 fueron muy superiores a lo normal: recargaron el acuífero haciendo subir su nivel en unos 10 metros, restaurando casi la situación inicial.

Por tanto, la naturaleza está realizando ya un sistema de almacenamiento a largo plazo, con un tiempo de residencia de unos 10 a 15 años, que representa la mayor seguridad contra las sequías o años de precipitación escasa. Sospecho que en algunos lugares se podría agotar un embalse durante un largo período, hasta de 40 años, y rellenarlo aun en ese caso mediante procesos naturales. Sabemos que la naturaleza llegó a tiempos de almacenamiento mucho más largos (siglos o milenios), en los enormes embalses de agua subterránea que existen bajo desiertos como el del Sáhara. Estudios recientes han demostrado que el agua recuperable de esos embalses permitiría el riego de varios cientos de miles de hectáreas de tierra durante siglos, aun suponiendo que la recarga natural de los embalses fuera insignificante. La perpetua búsqueda de más agua por parte del hombre significa, sin embargo, que debería dedicarse una inmediata atención a la recarga artificial de los embalses subterráneos.

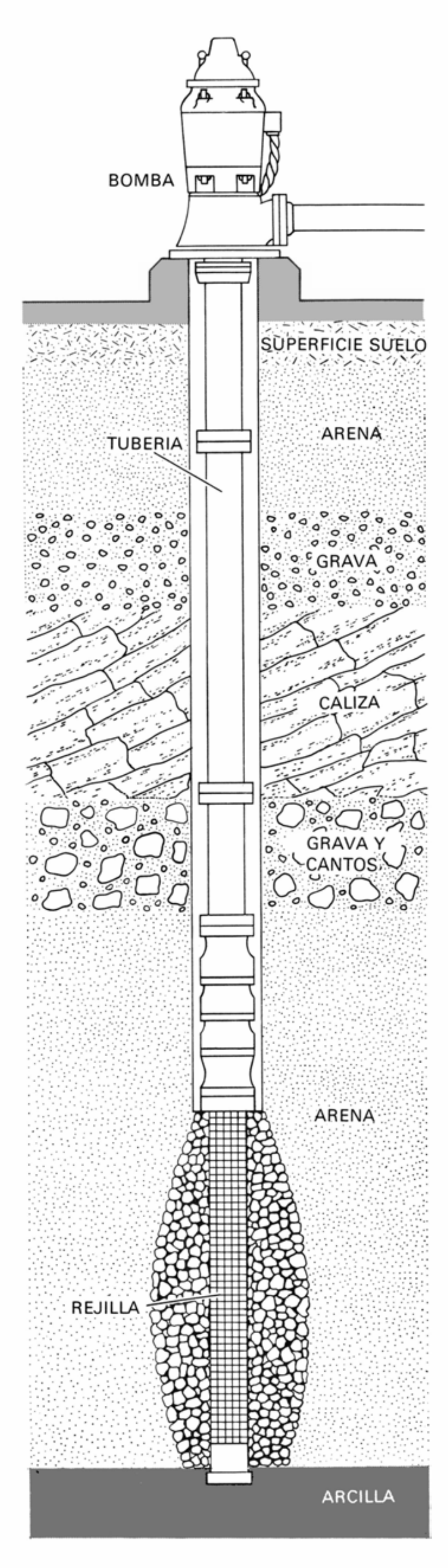
La práctica de la recarga artificial empezó en nuestro siglo. En general se hizo a pequeña escala y con objetivos limitados: restaurar el nivel de una superficie piezométrica que estuviera descendiendo rápidamente, o mejorar la calidad del agua. La mayor parte de la labor se hizo en California, donde unas 300 estaciones recargaron un total de unos ocho kilómetros cúbicos de agua entre 1900 y 1960. A partir de 1960, la cuantía de la recarga artificial ha llegado a ser de un kilómetro cúbico por año. También en Europa se ha practicado una recarga artificial de cierta magnitud, principalmente para mejorar la calidad del agua.

Los programas más recientes de California e Israel que he mencionado suponen una recarga artificial a mayor escala; implican asimismo el concepto de almacenamiento a largo plazo en embalses de agua subterránea. Sin embargo, el ejemplo de recarga artificial más llamativo se encuentra en el valle del Indo, aunque hasta hace unos 15 años era un proceso incontrolado y en gran medida incidental.

El proyecto del Indo es el mayor sistema de riego del mundo, y abarca más de 10 millones de hectáreas. Como he mencionado, se desarrolló en el sistema del río Indo durante el siglo pasado. Una red de 60.000 kilómetros de canales, en Pakistán, recubre la mayor parte de un vasto embalse de agua subterránea, que comprende 16 millones de hectáreas y se prolonga hasta 350 metros en profundidad. El embalse contiene una capa de fondo de agua relativamente salina (6000 partes por millón, comparada con las 35.000 partes por millón que corresponde al agua de mar típica) y una capa superior con agua de mejor calidad (700 partes por millón).

El embalse está siendo recargado continuamente por escapes procedentes de los canales de riego. Esta recarga artificial se calculó en 1965 en 17 kilómetros cúbicos por año, mientras que la recarga natural es de 8 kilómetros cúbicos. Como resultado de estas entradas, la superficie piezométrica se elevó en 30 centímetros por año, ocasionando lagunas de imbibición en muchas partes del valle del Indo y problemas de salinidad en el suelo.

Un grupo de la Universidad de Harvard estudió esta situación como parte de un proyecto patrocinado por el Banco Mundial. El equipo propuso usar el agua subterránea para riego y administración de los recursos hídricos. Parte del plan consistía en extraer agua durante 30 años para hacer descender la superficie piezométrica y reducir el problema



POZO HABILITADO, que en la región del Indo se denomina *pozo entubado.* Estos pozos constituyen el medio por el que se hace descender el agua subterránea en el valle del Indo, así como una fuente de riego y también de administración de los recursos de agua subterránea. El pozo se perfora hasta una profundidad de 30 a 70 metros, colocando en el agujero una tubería que acaba en un filtro que penetra hasta cierta profundidad por debajo de la superficie piezométrica. El filtro no permite la penetración de detríticos gruesos.

de la salinidad. El agua así bombeada debe diluirse con el agua de canal para mantener un nivel aceptable de salinidad cerca de las raíces de las plantas regadas. Además, el agua extraída del embalse subterráneo debe llevarse hasta el mar a través de los canales.

El agua atrapada del ciclo hidrológico por el sistema del Indo durante el siglo pasado, y almacenada en embalses subterráneos, representa actualmente un volumen de unos 400 kilómetros cúbicos de agua útil, guardada en los 30 metros superiores de la parte del acuífero sobre la que descansa el área regada. El agua se ha extraído por medio de un sistema de pozos entubados. La mayoría de ellos, perforados por la empresa privada, hasta una profundidad de unos 30 metros, extraen agua a razón de unos 30 litros por segundo. Los pozos entubados instalados por organismos estatales llegan a una profundidad de unos 70 metros y descargan a razón de unos 120 litros por segundo.

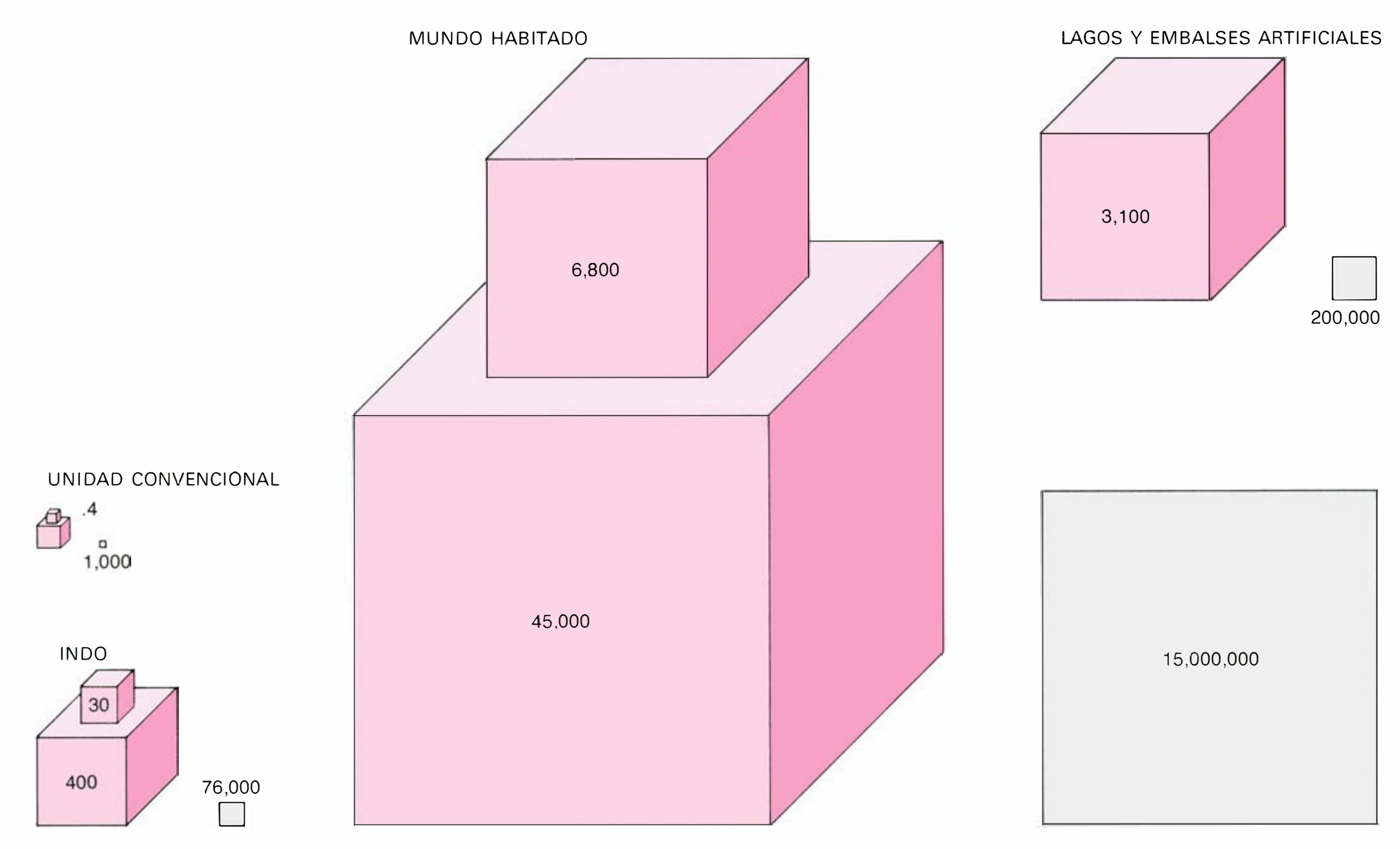
De estas actividades resulta que el agua subterránea, considerablemente complementada por la recarga artificial, está contribuyendo a un tercio del suministro de agua en la región y ha permitido aumentar en cuatro millones de hectáreas la extensión de terreno del que pueden recogerse cosechas. Cada hectárea de terreno regado podría beneficiarse también de 40.000 metros cúbicos de agua útil almacenada subterráneamente durante los últimos cien años. La explotación de esta agua almacenada podría permitir una ampliación anual del área cultivada, o una garantía de pleno riego en tiempos de grave sequía o de varios años de precipitación escasa. A los precios actuales, podría producirse, a partir del embalse de agua subterránea, un kilómetro cúbico de agua útil, por año, por unos 20 millones de dólares, mientras que la misma cantidad de agua retenida del ciclo mediante una presa y su embalse costaría 100 millones de dólares.

Tal sistema encierra la clave para el futuro de la agricultura y de la gestión del agua: que la producción de alimento pueda avanzar al mismo paso que el incremento de población y que la disponibilidad de agua aumente a la par que las crecientes demandas de recursos hídricos. La utilización conjunta de embalses superficiales y subterráneos aliviará las deficiencias de agua, estacionales y a largo plazo. A fin de que tal sistema funcione satisfactoriamente, se requiere que los embalses subterráneos estén bien equipados para una explotación que pudiese continuar durante cualquier periodo de tiempo: desde unos cuantos meses por año hasta varios años por década. Preparar un sistema de este tipo exige decisiones políticas que hasta ahora ni siquiera se han empezado a considerar en la mayoría de los cuerpos legales.

La aplicación de una gestión sistemática a los acuíferos se encuentra en la fase exploratoria. Nuevas experiencias podrían mejorar las técnicas empleadas en la administración de recursos hídricos en general y de los embalses de agua subterránea en particular.

La tecnología de la recarga artificial de embalses de agua subterránea constituye un caso ejemplar. Si esta tecnología se perfeccionase, los embalses de agua subterránea servirían tanto para almacenamiento estacional como prolongado. Además, aumentaría el suministro total de agua utilizable, dado que el sistema conservaría agua que, de otra forma, se evaporaría o escaparía sin utilizar.

La infiltración a partir de los canales de riego es una de las mejores técnicas para recargar artificialmente a gran escala los embalses de agua subterránea. Por consiguiente, el revestimiento de canales, que a menudo se efectúa para reducir las filtraciones, es de hecho un



ALMACENAMIENTO Y RECARGA de agua en los 30 metros superiores de los embalses de agua subterránea ilustrados según una unidad patrón, para la región del Indo y para el mundo habitado. La cifra al lado de un cuadrado representa, en kilómetros cuadrados, el terreno suprayecente a embalses de aguas subterráneas; el número del cubo superior es la de recarga anual, incluida la artificial, en el Indo, en kilómetros cúbicos; el número del cubo inferior muestra de cuántos kilómetros cúbicos adicionales de agua útil y recuperable se dispone. Se muestra también, como término de comparación, la magnitud del área superficial y de la recarga anual en lagos y embalses artificiales.

gasto inútil. Sería mejor remodelar los canales y agrandarlos, con la condición de que el trabajo estuviera estrechamente coordinado con la instalación de piezómetros, para regular el ascenso de la superficie piezométrica (water table).

En zonas templadas, otro medio de recarga de embalses de agua subterránea consiste en derramar las aguas de crecida de los deshielos invernales sobre terreno irrigable pelado; esto es, regar el terreno aunque no se esté cultivando nada en él. La técnica proporcionaría un almacenamiento ocasional en el suelo que sería útil para las cosechas cuando éstas se plantaran y un almacenamiento a largo plazo en el embalse de agua subterránea subyacente a la zona regada.

Una técnica simple y muy antigua para la recarga de los embalses de agua subterránea consiste en construir lagos y pequeñas presas en las cuencas altas de los ríos, de manera que parte del agua que normalmente se escurriría, se infiltre en el terreno. Entre otras técnicas, más caras, está la de construir hoyos de infiltración (que es un tipo de excavación destinada a recoger agua de la escorrentía superficial de tormentas, que, de otra forma, correría inútilmente hacia el mar) y pozos de recarga, por medio de los cuales se inyecta agua en el terreno, en vez de extraerla de él. En Estados Unidos se han instalado numerosos hoyos de infiltración y pozos de recarga en Long Island, Arkansas, los Grandes Llanos del norte y algunos estados del oeste.

A escala global, el problema con que se enfrenta la humanidad no es la falta de agua dulce, sino la carencia de regímenes eficientes para usar el agua disponible. La necesidad de producir más alimentos continuará y solamente podrá ser satisfecha mediante un abastecimiento garantizado de agua en las zonas agrícolas. El déficit de 400 kilómetros cúbicos de agua que originó la crisis de alimentos de 1972 es insignificante si se considera que aún hay unos 45.000 kilómetros cúbicos de agua almacenados en los 30 metros superiores de los embalses de agua subterránea.

La humanidad puede resolver el problema del agua si los gobiernos y quienes están encargados de adoptar decisiones se hallan dispuestos a acometer una gestión apropiada de sus recursos hídricos y, en especial, de sus reservas subterráneas. Además, desde el punto de vista técnico, se está consiguiendo hacer frente a acontecimientos adversos tales como las deficiencias de agua provenientes de cambios en el comportamiento del tiempo, iniciando la regulación a gran escala y a largo plazo del ciclo del agua a través de los embalses de agua subterránea.

Ramapithecus

Este primate, ya extinguido, es el primer homínido con características humanas bien definidas. El hallazgo de nuevos restos ha permitido delimitar su situación en la filogenia humana

Elwyn L. Simons

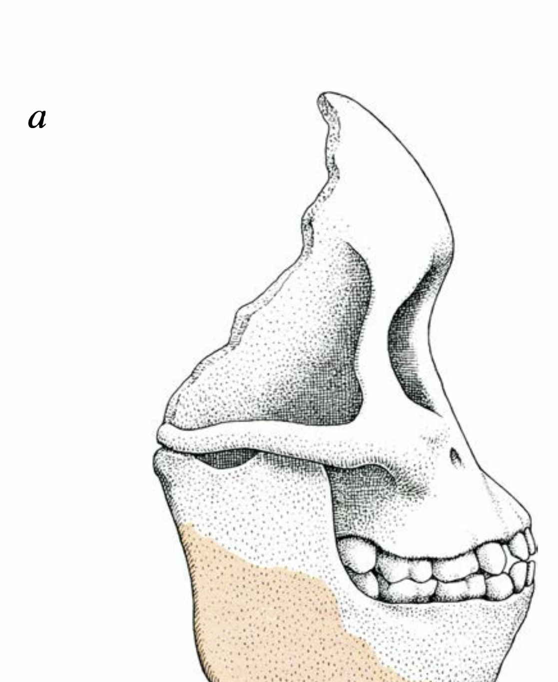
Hace quince años, el único dato existente sobre la aparición de una rama definitivamente humana en la genealogía de nuestra especie era un fragmento de mandíbula superior que había sido descubierto en el norte de la India en 1932. Hoy día, se han encontrado otros fósiles de este homínido y de otros géneros próximos en Africa oriental, Grecia, Turquía, Hungría, la India y Pakistán. Durante los años 60, estos fósiles de homínidos aumentaron considerablemente, no gracias a nuevos descubrimientos originales, sino como resultado de la identificación de restos que se encontraban en colecciones de fósiles de primates en diversos museos. Desde 1972, el número de estos restos casi se ha duplicado. Estos hallazgos han constituido un avance significativo para el paleontólogo especializado en primates. Desde el punto de vista de la evolución de la especie humana, estos hallazgos han permitido delimitar con mayor exactitud los cambios evolutivos que se han producido durante los últimos 14 millones de años. Estos estudios posibilitan, en la actualidad, definir el camino que ha llevado de los hominoides (rama principal del árbol genealógico humano, en la que se encuentran situados los grandes antropoides) a los homínidos y de los homínidos al género *Homo*.

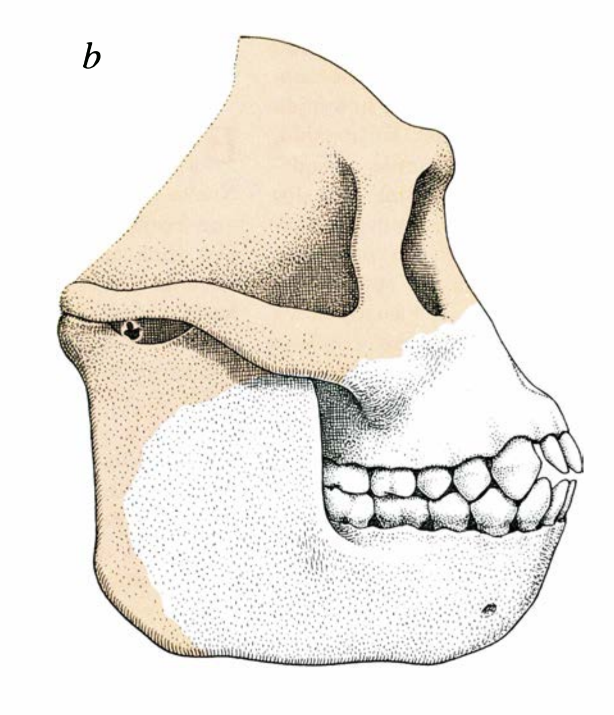
Este camino se inicia en el Mioceno, momento en que vivía en el Viejo Mundo una población de antropoides cuya existencia se conoce desde hace más de un siglo. En 1856, Edouard Lartet, abogado y paleontólogo francés, describió una mandíbula de primate hallada en un terreno arcilloso del Mioceno en los Pirineos franceses. Lartet denominó a esta especie fósil *Dryopithecus fontani*. El nombre genérico, que combina los términos griegos que significan "roble" y "antropoide", pone de manifiesto la idea de Lartet, quien pensaba que este primate vivía en el bosque. Esta conjetura se apoya en los restos de plantas y de animales encontrados junto a otros fósiles de *Dryopithecus*. Estos antropoides cosmopolitas vivían, al parecer, en regiones boscosas tropicales y subtropicales, alimentándose de hojas y de frutas.

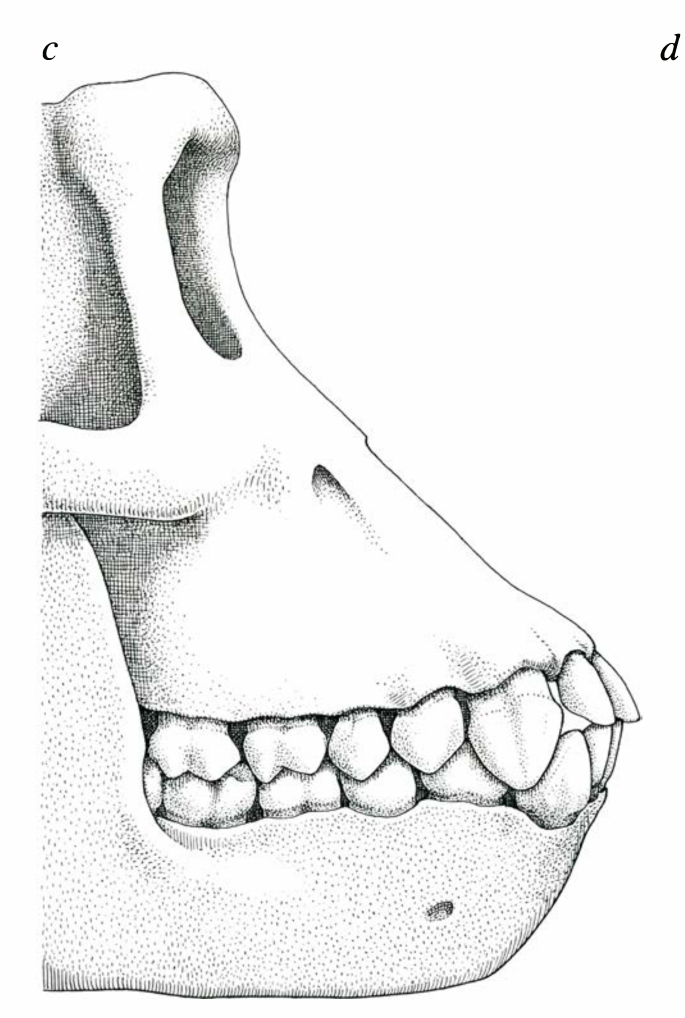
Los primeros fósiles de *Dryopithecus* hallados en Francia fueron tres restos parciales de mandíbula inferior, una de las cuales había perdido tan sólo un diente. No pudo encontrarse ninguna mandíbula superior, y, en general, la mayoría de los fósiles europeos de *Dryopithecus* no fueron más que dientes sueltos. En consecuencia, apenas existían datos acerca de la forma del cráneo, de la cara o del resto del cuerpo de este antropoide hasta finales de la década de los 40. Por esta época, L. S. B. Leakey y colaboradores hallaron restos más completos en los ricos yacimientos del Mioceno situados en las islas del lago Victoria y al noreste del mismo. Entre estos restos se encontraron fragmentos de mandíbula de tamaño parecido al de los actuales gibones (el más pequeño de los antropoides actuales) y al de los gorilas. En 1948, la esposa de Leakey, Mary, encontró un cráneo fósil perfectamente conservado en la isla de Rusinga del lago Victoria, y, tres años más tarde, Tom Whitworth, colaborador de Leakey, descubrió restos de un segundo cráneo en el mismo lugar, así como un miembro anterior y otros restos óseos, entre los que se encontraban una mano, otros huesos de las extremidades y parte de un pie. Estos antropoides fósiles africanos quedaron clasificados dentro del género *Proconsul;* se identificó a los dos cráneos y los huesos de las extremidades como pertenecientes a la especie *Proconsul africanus*. Sin embargo, estudios recientes sostienen que *Proconsul* no constituye un género aparte, sino que se trata de un miembro africano del género *Dryopithecus*.

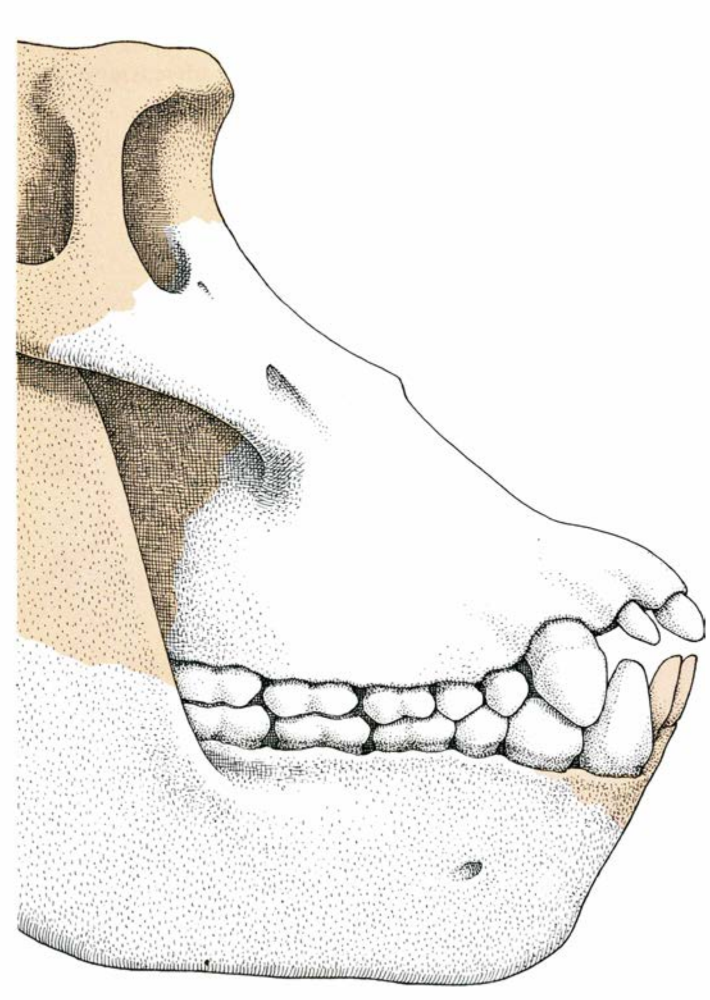
Antes de seguir hay que señalar otros dos datos importantes con respecto a los dryopithecinos. Uno de ellos es que se encuentran fósiles de dryopithecinos tanto en yacimientos del Mioceno como en estratos del período anterior u Oligoceno. Ello indica que los dryopithecinos vivieron durante un período de unos 20 millones de años. El otro dato importante es que estos fósiles han sido hallados en Francia, en Africa oriental y en otros lugares correspondientes a una amplia región del mundo antiguo: el desierto occidental de Egipto, la zona de Barcelona, el valle del Rhin, la zona de Viena, las montañas del noreste de Hungría, la provincia de Macedonia (Grecia), Asia Menor, las llanuras de Potwar (Pakistán), las colinas de Siwalik (la India), las minas de carbón de Yunán (China occidental) y diversos lugares de la zona central del sur de China. Durante este extenso y largo proceso de irradiación, estos antropoides se encontraron con climas cada vez más fríos, en los que los bosques tropicales o subtropicales daban paso a bosques abiertos o a sabanas. Los fósiles hallados en Europa y Asia indican que, hace unos 15 millones de años, los *Dryopithecus* produjeron por lo menos otros tres géneros.

Dos géneros de ellos, *Sivapithecus* y *Gigantopithecus*, eran primates con un macizo facial tan alargado como el de los antropoides actuales. El tercer género, *Ramapithecus*, presentaba un macizo facial de pequeño tamaño. Las relaciones existentes entre estos tres primates del Mioceno no podrán ser determinadas hasta que se encuentren restos que comprendan bóvedas craneales y huesos de las extremidades. En ge-







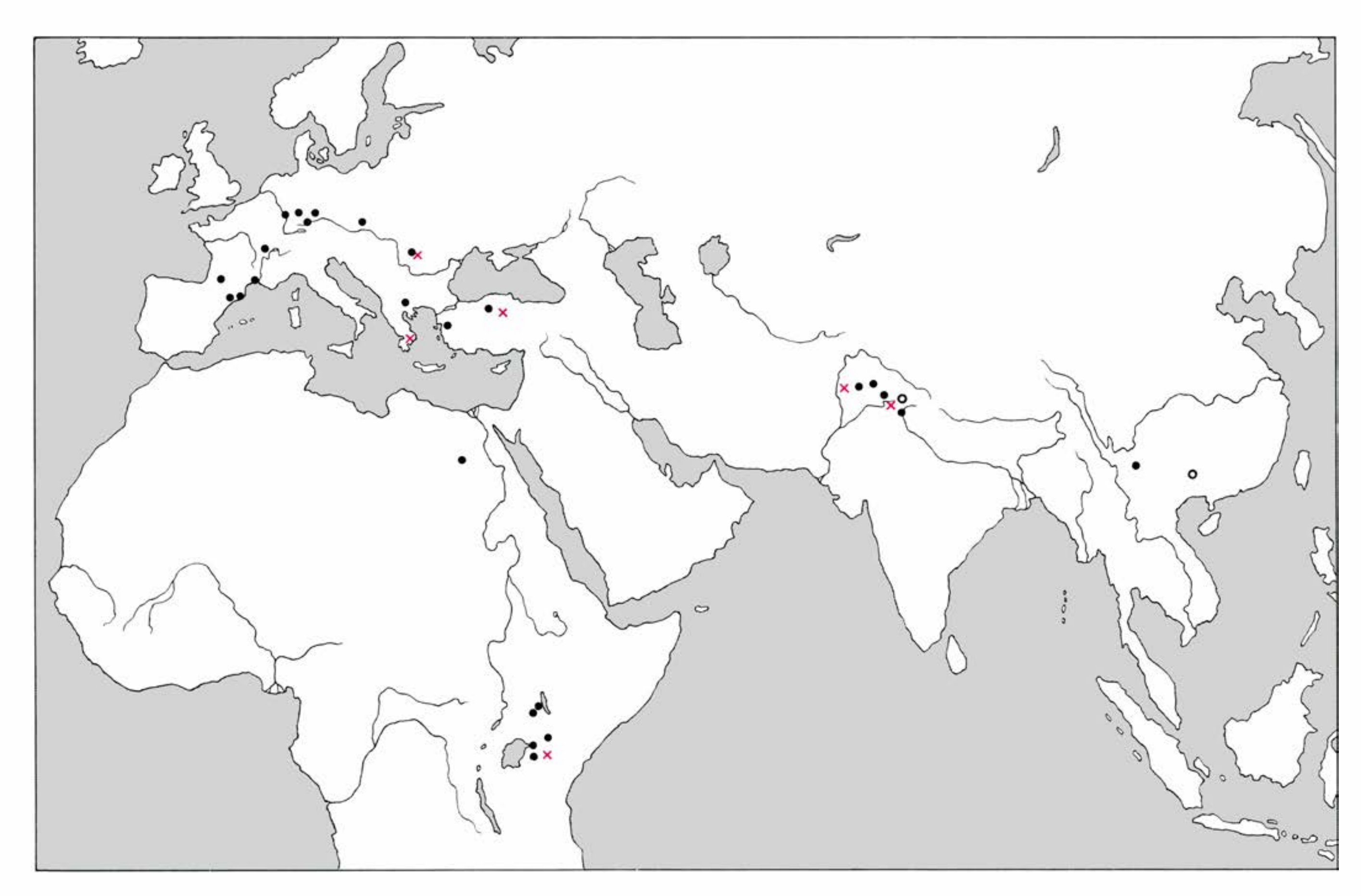


LOS PERFILES DE CUATRO PRIMATES permiten comparar el tamaño y estructura de sus macizos faciales. El primero corresponde al individuo juvenil de Taung *(a)*, resto de *Australopithecus* descrito por Raymon Dart en 1925. El segundo corresponde a una reconstrucción de *Ramapithecus (b)*, homínido de tamaño parecido al *Australopithecus* juvenil. El perfil se basa en los datos obtenidos de restos recientemente descubiertos *(las zonas coloreadas son aún especulativas)*. El tercero presenta el aspecto de un gorila hembra actual *(c)* y el cuarto es reconstrucción de un antropoide de gran tamaño del Mioceno, *Sivapithecus (d)*, contemporáneo de *Ramapithecus*. Los cambios que han sufrido la mandíbula superior y la mandíbula inferior, como consecuencia de diferencias en la dieta de unos y otros, han dado lugar a que el homínido citado presente un menor prognatismo que los antropoides fósiles o vivientes, lo que hace que su aspecto sea más parecido al que deben haber tenido otros homínidos, los *Australopithecus*, juveniles o adultos *(véanse las ilustraciones de las páginas 16 y 17)*.

neral, *Sivapithecus* ha sido clasificado como un dryopithecino y, en consecuencia, considerado no como un homínido sino como un antropoide. En cambio, *Ramapithecus* ha sido siempre considerado como miembro del linaje de los homínidos. Sin embargo, tanto *Sivapithecus* como *Gigantopithecus* presentan ciertas características que son típicas de los homínidos. Así, por ejemplo, los tres géneros poseen molares con una gruesa capa de esmalte. Estas semejanzas se deben probablemente a que los tres géneros se hallan relacionados entre sí, aunque también podrían ser debidas a respuestas frente a cambios ambientales parecidos. De los tres géneros, el que presenta un mayor parecido con los homínidos es *Ramapithecus*. Sin embargo, el hallazgo de nuevos fósiles, tanto de *Sivapithecus* y de *Gigantopithecus* como de *Ramapithecus*, obligará, sin duda, a cambiar los conceptos existentes en la actualidad acerca de la forma en que los animales de tipo humano se separaron de los antepasados de los antropoides. A continuación, vamos a referirnos a una de las últimas ramas de la genealogía humana, tomando por punto de partida a *Ramapithecus*.

En las colinas de Siwalik, en la India, a unos 150 kilómetros al norte de Nueva Delhi, existe un grupo de aldeas que recibe el nombre de Haritalyangar, zona en la que se encuentran yacimientos fósiles del Mioceno que llegan hasta Pakistán. En esta zona, G. Edward Lewis, joven estudiante postgraduado de la Universidad de Yale, que había alcanzado la región en solitario, viajando a caballo, descubrió, en 1932, una serie de fósiles a los que identificó como "antropoides de aspecto humano". Uno de ellos, del que obtuvo una mandíbula superior, fue clasificado por Lewis como perteneciente a un nuevo género y especie, a los que denominó *Ramapithecus brevirrostris*. El nombre genérico significa antropoide de Rama, siendo Rama el héroe de un poema épico de la India. El nombre específico resulta más descriptivo, ya que procede de las palabras latinas que significan "cara corta", característica que permite diferenciarlo de los antropoides. Lewis señaló el aspecto humano de la mandíbula y de los dientes que contenía. Al escribir su tesis doctoral, en 1937, dicho autor situó al nuevo género en la familia Hominidae, grupo del orden Primates cuyo único representante actual es *Homo sapiens*.

Hasta 1961 no se encontró ningún otro fósil de *Ramapithecus*. El nuevo hallazgo, procedente de Fort Ternan, en el sudoeste de Kenya, se debió a L. S. B. Leakey, e incluía fragmentos de ambas ramas de una mandíbula superior. Leakey, con su habitual autosuficiencia, describió estos restos como si correspondiesen a un género totalmente nuevo, al que dio el nombre de *Kenyapithecus wickeri* en honor de su amigo, Fred Wicker, en cuya explotación encontró el fósil. *A posteriori*, es probable que si las mandíbulas superiores de Haritalyangar o de Fort Ternan se hubiesen acompañado de las correspondientes mandíbulas inferiores, o si Leakey hubiese comparado su hallazgo con los restos de Haritalyangar, que se encuentran en Yale, el problema de nomenclatura habría sido mucho menor.



PRIMATES MAS EVOLUCIONADOS del Oligoceno, del Mioceno e incluso de períodos posteriores. Entre ellos existía un grupo de antropoides muy abundante y cosmopolita: los dryopithecinos *(puntos negros)*. **La mayoría de ellos corresponden a dos géneros:** *Dryopithecus*, **distribuido principalmente en Africa y Europa, y** *Sivapithecus*, **distribuido sobre todo al este de Asia Menor. Estos antropoides existieron hace más de 25 millones de años. Tales grupos dieron origen al antropoide** *Gigantopithecus (círculos negros)* **y a ciertos homínidos como** *Ramapithecus*, **género distribuido por lo menos en cinco regiones de Eurasia y en una de Africa Oriental** *(cruces en color)*. **Al igual que** *Ramapithecus*, *Gigantopithecus* **presenta modificaciones anatómicas de las mandíbulas y de los dientes, que reflejan cambios dietéticos.**

Sin embargo, esta conclusión es excesivamente simplista, ya que otro resto de *Ramapithecus,* hallado en Grecia durante la Segunda Guerra Mundial, y descrito en 1972, fue también clasificado como perteneciente a un nuevo género y a una nueva especie: *Graecopithecus freyburgi.* El nombre específico se eligió en honor de Bruno von Freyburg, geólogo alemán que encontró el fósil en un yacimiento miocénico mientras se hallaba destinado en Atenas durante la ocupación alemana de Grecia. Von Freyburg encontró la región dental completa de una mandíbula inferior con la totalidad de los dientes, aunque algunos de ellos desaparecieron posteriormente. Sin embargo, el fósil fue dejado de lado, ya que otro especialista lo identificó, equivocadamente, como una mandíbula de *Mesopithecus,* mono del Mioceno extraordinariamente abundante. Antes de que la mandíbula fuese siquiera fotografiada, se perdieron la mayoría de los dientes, que desaparecieron durante uno de los bombardeos de Berlín.

El siguiente fósil de *Ramapithecus* fue una mandíbula inferior, descubierta en un depósito del Mioceno cercano a Çandir, a unos 60 kilómetros al noreste de Ankara, en Turquía. El fósil fue hallado en 1973, y descrito aproximadamente un año después. En la descripción, este fósil fue clasificado como perteneciente a otro género, y a una especie totalmente nueva: *Sivapithecus alpani. Sivapithecus,* o antropoide de Siva es el nombre que Lewis había dado a uno de los restos que encontró en la India: el nombre específico se eligió en honor del director del Servicio Geológico de Turquía.

En los depósitos carboníferos del Mioceno, situados en los montes Rudabanya, al nordeste de Hungría, se descubrieron también diversos fósiles parecidos a los de *Ramapithecus.* Estos restos, hallados a lo largo de varios años, se describieron en 1975. También en este caso se utilizó una nueva denominación genérica y específica: *Rudapithecus hungaricus.*

Entre los fósiles más recientes de *Ramapithecus,* tan sólo los descubiertos en Pakistán, durante los últimos 18 meses, por mi colega de Yale David R. Pilbeam, se asignaron inmediatamente a *Ramapithecus.* Hasta un fragmento de mandíbula inferior procedente de Fort Ternan y clasificado en 1971 como correspondiente a *Ramapithecus* por Peter Andrews, del British Museum (Natural History), había sido previamente adjudicada a *Dryopithecus.*

Este problema de nomenclatura no se debió exclusivamente a la tendencia de los especialistas en taxonomía a plantear nuevas subdivisiones (en lugar de intentar sintetizar). Así, Ibrahim Tekkaya, paleontólogo turco que clasificó la mandíbula de Çandir como correspondiente a *Sivapithecus,* apenas disponía de otro material con el que comparar su fósil. La información más importante de que disponía era la disputa a propósito de una mandíbula de *Dryopithecus* que había sido asignada a *Ramapithecus* por error, en 1938. El error había venido repitiéndose en la literatura científica. Tekkaya llegó a la conclusión de que las diferencias entre su mandíbula y la que erróneamente se adjudicaba a *Ramapithecus* hacían imposible clasificar su hallazgo dentro de dicho género. Posteriormente, Tekkaya y Andrews publicaron un trabajo conjunto en el que se reparó el error y clasificaron la mandíbula de Çandir dentro del género *Ramapithecus.*

CATEGORIAS SUPERIORES			GENEROS
SUPERFAMILIA HOMINOIDEA (TODOS LOS ANTROPOIDES Y TODOS LOS HUMANOS Y PREHUMANOS)	**FAMILIA PONGIDAE** (TODOS LOS GRANDES ANTROPOIDES)	GRANDES ANTROPOIDES VIVOS	*PONGO* (ORANGUTAN) *PAN* (CHIMPANCE) *GORILLA* (GORILA)
		GRANDES ANTROPOIDES EXTINGUIDOS	*DRYOPITHECUS* *SIVAPITHECUS* *GIGANTOPITHECUS*
	FAMILIA HOMINIDAE (TODOS LOS HUMANOS Y PREHUMANOS)	HUMANOS VIVOS	*HOMO* (HOMBRE)
		PREHUMANOS EXTINGUIDOS	*RAMAPITHECUS* *AUSTRALOPITHECUS*

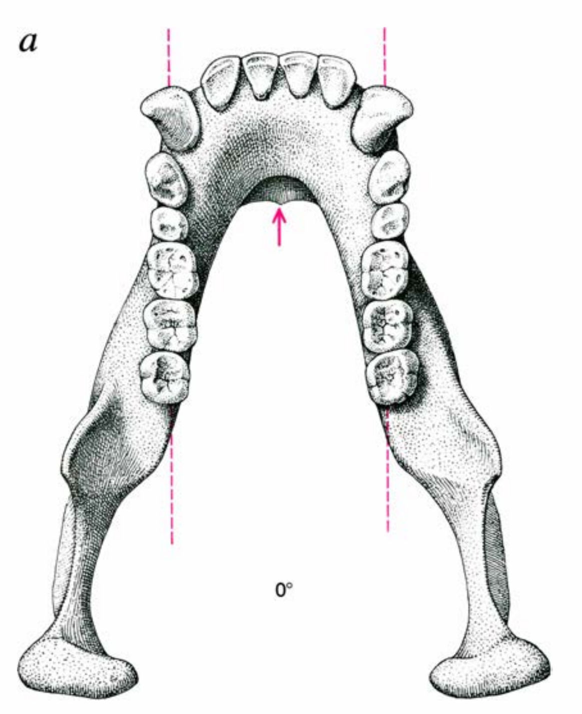
LA SUPERFAMILIA HOMINOIDEA, que forma parte del orden Primates, comprende tres familias. La familia Pongidae abarca, además de los tres géneros de antropoides actuales, tres géneros fósiles; la familia Hominidae, al género *Homo* y a dos géneros prehumanos extinguidos.

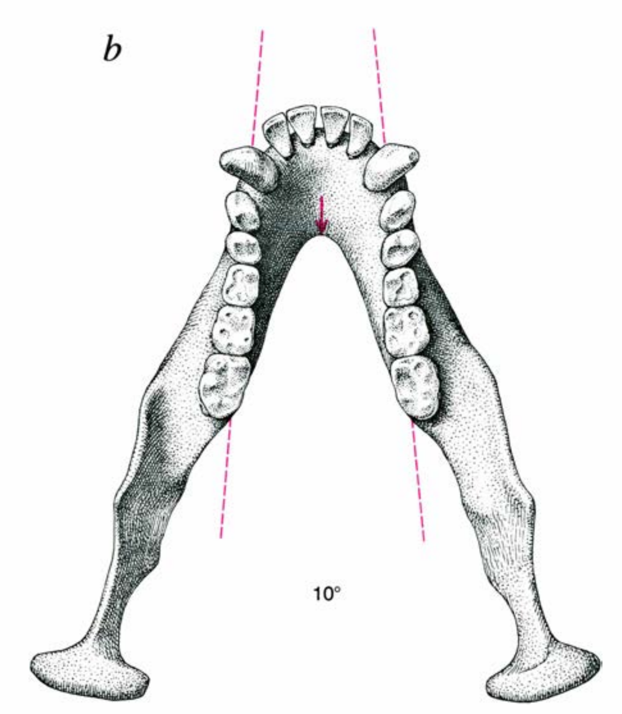
Igualmente, el paleontólogo húngaro Miklos Kretzoi, que clasificó sus descubrimientos como correspondientes a *Rudapithecus,* no hubiera cometido probablemente ese error si la primera mandíbula inferior hallada en Hungría, en 1969, no se hubiese encontrado en un estado deplorable. Si se comparan los restos húngaros con los fósiles de Rudabanya, que comprenden mandíbulas superiores e inferiores, el parecido con *Ramapithecus* resulta inconfundible. En la actualidad, Kretzoi reconoce que sus fósiles son de tipo *Ramapithecus,* aunque no se correspondan exactamente.

Por último, el eminente paleoantropólogo G. H. R. von Koenigswald, que clasificó la mandíbula inferior hallada en Atenas como correspondiente a un nuevo género, *Graecopithecus,* debido a que en la época no se conocían mandíbulas inferiores de *Ramapithecus* completas, las primeras de las cuales fueron descritas en 1974 (mandíbula de Çandir) no podía, ni siquiera en 1972, haberse dado cuenta del parecido existente entre el fósil encontrado en Atenas y los procedentes de Fort Ternan y Çandir.

Aparte de los citados, existen otros restos de *Ramapithecus* que han tenido que ser clasificados de nuevo. Así, por ejemplo, en 1965, Pilbeam y el autor aceptaron que los homínidos no poseían la cresta "simiesca", característica de monos y antropoides, siguiendo las ideas más en boga en dicho momento. La cresta simiesca es un torus, o eminencia ósea horizontal, situada en la cara interna de la mandíbula inferior. Al escribir el artículo citado, consideraron que su presencia en una mitad de mandíbula inferior hallada cerca de Domeli, en Pakistán, no permitía clasificar a este resto fósil dentro del género *Ramapithecus.* Posteriormente, llegamos a la conclusión de que tanto *Ramapithecus* como los homínidos africanos de posterior aparición, conocidos como *Australopithecus,* presentaban crestas óseas en la cara interna de la mandíbula inferior. En 1969, Pilbeam recuperó para el mundo científico la mandíbula de Domeli, que había sido injustamente olvidada, describió sus características de tipo homínido y la clasificó entre los restos correspondientes a *Ramapithecus.*

En el momento en que escribo este artículo, se han descrito unas 24 mandíbulas de *Ramapithecus.* ¿Qué es lo que estos restos fósiles nos dicen acerca del más remoto antepasado de la especie humana? Los datos existentes en la actualidad son principalmente de tipo anatómico, pero conducen a una conclusión inevitable: *Ramapithecus* se había adaptado a una forma de vida muy distinta de la de sus cercanos parientes, los *Dryopithecus,* que vivían en los bosques. Sus mandíbulas y dientes muestran adaptaciones totalmente comparables con las del homínido africano *Australopithecus.* Sin embargo, entre uno





ESTAS CUATRO MANDIBULAS INFERIORES permiten determinar la presencia de diferencias en el grado de divergencia posterior de los arcos dentales en tres primates fósiles. A la izquierda *(a)*, y con finalidades comparativas, se incluye la mandíbula de un chimpancé actual; el arco dental, con típica forma en *U*, se halla formado por ramas paralelas; en consecuencia, su grado de divergencia es nulo. A continuación *(b)* aparece la reconstrucción de la mandíbula de *Dryopithecus;* las ramas mandibulares presentan una divergencia *(en color)* de unos 10 grados. Inmediatamente después *(c)* se encuentra una reconstrucción de la mandíbula de *Ramapithecus.* Las ramas mandi-

y otro existe un amplio hiato. Las especies de *Ramapithecus* vivieron en Eurasia hace por lo menos ocho millones de años, mientras que los primeros fósiles de *Australopithecus* y *Homo,* hallados por Donald C. Johanson, de la Case Western Reserve University, en Etiopía, y, por Mary Leakey, en Tanzania, no llegan a los 4 millones de años.

Volviendo a las características anatómicas, en 1967, al estudiar de nuevo la morfología mandibular de los antropoides fósiles mejor conocidos, advertí, con sorpresa, que los antropoides del Mioceno presentaban una arcada dental (el arco formado por los dientes, visto por la norma superior) muy distinta de la que se observa en los antropoides actuales. En los antropoides actuales, este arco presenta forma de *U,* ya que las ramas mandibulares son paralelas. En cambio, en los antropoides del Mioceno dicho arco presenta forma de *V,* como resultado de la divergencia posterior de los arcos mandibulares.

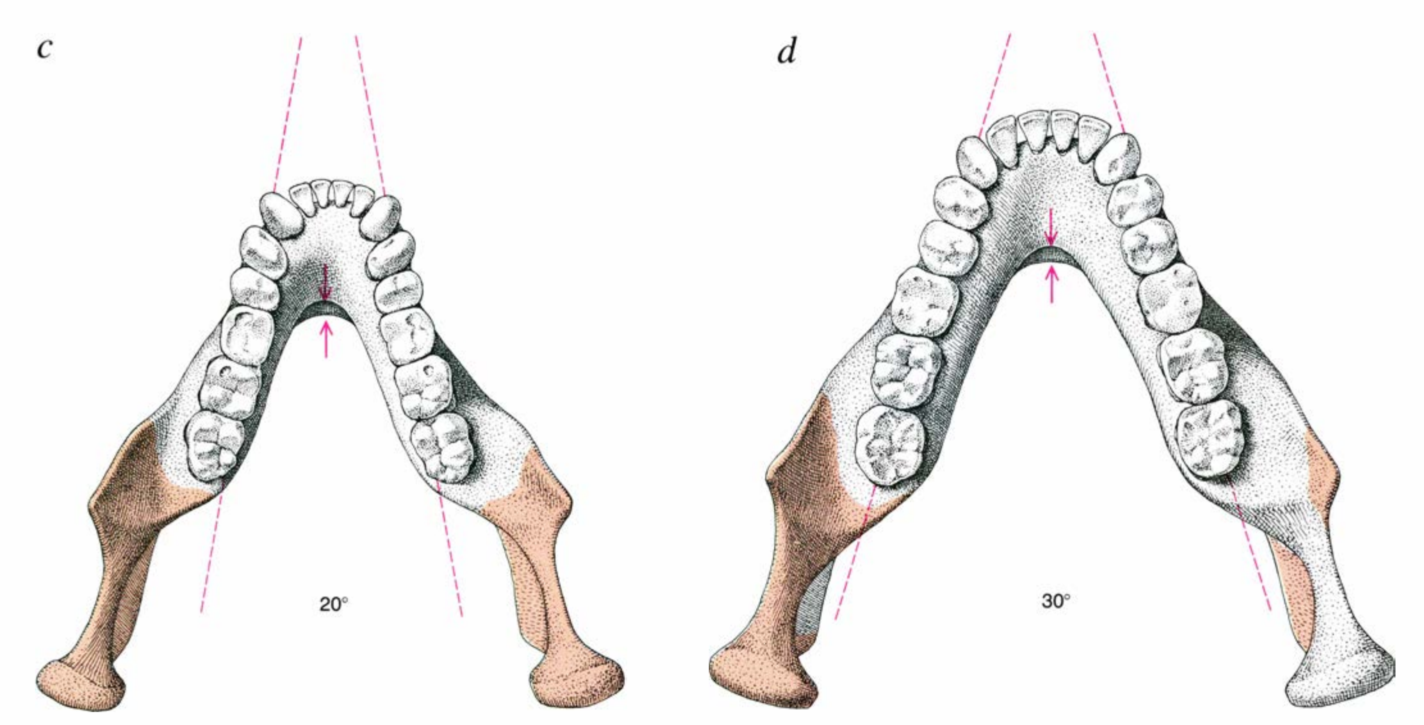
Dos años después, E. Genet-Varcin, en París, observó que el arco dental de la mayoría de los *Australopithecus* presentaba también forma de *V.* En consecuencia, dos grupos separados entre sí por varios millones de años poseen una característica idéntica (en ambos grupos, la *V* es de vértice romo). En cambio, el arco dental de *Homo sapiens* no tiene forma ni de *U* ni de *V,* sino que es redondeado. En consecuencia, es probable que el arco en forma de *U* de los antropoides actuales y el arco semicircular humano procedan del arco original en forma de *V.*

¿Cuáles son las características del arco dental de *Ramapithecus*? Básicamente, el arco dental de *Ramapithecus* es de tipo intermedio. Las ramas del maxilar divergen, formando con las líneas paralelas un ángulo de unos 20 grados. Este ángulo es superior al que presentan los antropoides del Mioceno (unos 10 grados) y menor que el de los *Australopithecus* (unos 30 grados).

Tanto *Ramapithecus* como *Australopithecus* poseyeron mandíbulas cuyo grosor desde la cara lingual a la cara correspondiente a la mejilla es comparable, a nivel de la región molar, con la altura de la mandíbula. En los antropoides actuales, la mandíbula es mucho menos robusta, siendo su grosor inferior a su altura. Los molares de *Ramapithecus* y de *Australopithecus* presentan también un aspecto muy robusto. Se trata de molares planos, de gran tamaño, y, vistos desde arriba, ofrecen un aspecto redondeado. A diferencia de lo que ocurre en los molares de los chimpancés y de los gorilas, y también en los molares de *Dryopithecus* en sentido estricto (excluyendo a *Sivapithecus*), el esmalte es de considerable grosor y, en consecuencia, más resistente al desgaste. Los caninos e incisivos son de menor tamaño que los de los antropoides actuales. Además, la mandíbula inferior de *Ramapithecus* y de *Australopithecus* se encuentra mejor preparada para la masticación, al presentar no una, sino dos crestas internas. Estas eminencias, denominadas torus superior e inferior, se observan todavía en los fósiles de *Australopithecus* de hace tan sólo un millón de años.

Otra característica especial que presentan ambos homínidos es una marcada diferencia en el desgaste de los dos primeros molares. En *Dryopithecus* y en los antropoides actuales, el esmalte de las cúspides del primer molar se desgasta, apareciendo en su lugar el cemento subyacente, cosa que ocurre también en el segundo molar. En *Ramapithecus* y *Australopithecus* el distinto desgaste que presentan estas dos piezas indica que la erupción del segundo molar debía producirse mucho más tarde que la del primero, lo que implica que su maduración era mucho más lenta que la de los antropoides del Mioceno o de los actuales.

Este conjunto de características –mandíbulas gruesas y preparadas para la masticación, molares aplanados y reducción de los caninos y de los incisivos– puede compararse con las que se observan en ciertos tipos de mamíferos, como los herbívoros provistos de pezuñas, los roedores herbívoros o incluso los elefantes. Estos mamíferos mastican los alimentos triturándolos mediante un movimiento de balancín, lo que implica que la masticación es, por lo menos en



bulares encontradas presentan una divergencia de unos 20 grados. Por último *(d)* aparece la reconstrucción de la mandíbula de *Australopithecus.* La divergencia de las ramas mandibulares es de 30 grados. En homínidos posteriores, las ramas mandibulares presentan una divergencia aun mayor. Las flechas de color señalan las diferencias observadas en las protrusiones del borde mandibular que reciben el nombre de torus superior e inferior. Los antropoides actuales presentan un torus inferior muy pronunciado; en *Dryopithecus,* el torus superior era especialmente marcado. Ambas crestas se hallan bien desarrolladas en *Ramapithecus* y *Australopithecus* *(véase la ilustración de la página siguiente).*

parte, de tipo lateral. En cambio, si la masticación es de tipo cortante, las fuerzas que actúan sobre la mandíbula son de tipo vertical. En consecuencia, la trituración de alimentos implica que ciertos movimientos den lugar a fuerzas de tipo casi totalmente horizontal. Otra adaptación que presentan estos mamíferos consiste en el aumento de la longitud de la rama vertical de la mandíbula inferior, que forma un ángulo recto con la horizontal. En cambio, en el hombre moderno y en los chimpancés, dicha rama es de escasa longitud y forma un ángulo obtuso con la horizontal. En el hombre y en el chimpancé, el esmalte es delgado, y la mandíbula mucho menos robusta.

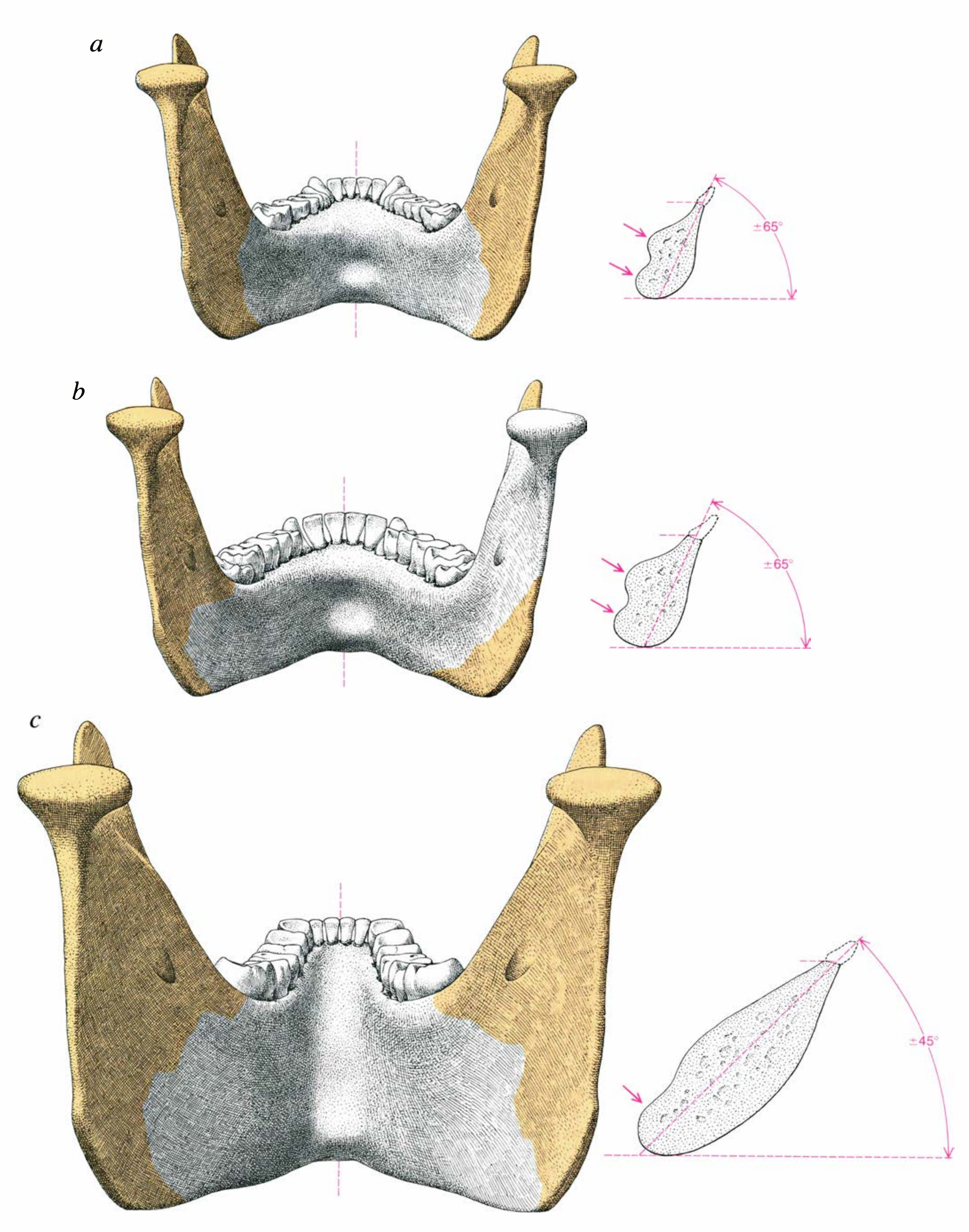
En consecuencia, no es de extrañar que la rama ascendente de la mandíbula de *Ramapithecus* y de *Australopithecus* forme con la horizontal un ángulo más recto que la del hombre o la de los antropoides africanos actuales. En *Ramapithecus* y *Australopithecus,* la rama ascendente es también de mayor longitud. La región anterior de dicha rama, que forma el proceso coronoide, se halla ligeramente dirigida hacia delante en relación al eje de los dientes. Este ángulo y desplazamiento se deben a las características de la mandíbula superior y de la porción inferior del macizo facial. La mandíbula superior posee alveolos dentales más profundos y gruesos; el hocico es corto y plano: las fosas nasales se hallan situadas sobre los incisivos superiores. Los pómulos se encuentran más adelantados, y los incisivos y caninos superiores no se dirigen hacia delante, sino directamente hacia abajo. La presencia de características tan parecidas entre estos dos géneros, distanciados por un largo período de tiempo, debe tener, necesariamente, algún significado.

Los primeros signos de cambio en cualquier adaptación importante de un organismo, que se ponen de manifiesto a través de cambios en la función de un sistema anatómico preexistente, permiten interpretar con mayor exactitud la historia de los fósiles. Estos cambios anatómicos son, con frecuencia, la antesala de importantes acontecimientos evolutivos. En el proceso de la evolución de los primates, el cambio adaptativo que quizá revista mayor importancia es el que dio lugar a la separación entre los antepasados de la especie humana y los que originaron los antropoides actuales, cambio que quizá pueda interpretarse a través de las adaptaciones funcionales de la dentadura de *Ramapithecus.*

¿En qué momento y por qué se separó *Ramapithecus* de los dryopithecinos, sus parientes más cercanos, a través de una anatomía mandibular predecesora de la que presentan los dos géneros posteriores de homínidos, *Australopithecus* y *Homo*? Para empezar, debemos determinar en qué lugares encontró *Dryopithecus* circunstancias ambientales que pudiesen favorecer la aparición de dicho cambio.

Entre los dryopithecinos, el grupo más íntimamente relacionado con *Ramapithecus* parece ser *Sivapithecus,* distribuido casi exclusivamente en Asia; los dryopithecinos africanos son bastante anteriores. Muchos de los fósiles antropoides más recientes descubiertos en Eurasia alcanzan gran tamaño, con mandíbulas muy robustas y esmalte muy grueso: todos ellos parecen corresponder al género *Sivapithecus.* Dos de estos fósiles de Eurasia han sido ya bien estudiados. Uno de ellos, *Sivapithecus indicus* se halla representado principalmente por restos procedentes del norte de la India, Cachemira, Pakistán y Turquía. El otro, que probablemente pertenece a un nuevo género, aún no descrito, procede del norte de Grecia, y recibió el nombre de *"Dryopithecus" macedoniensis,* en 1974. Después de su descubrimiento, se hallaron en el mismo lugar otros restos de mandíbulas de la misma especie; todas ellas se parecen a *Sivapithecus* más que a *Dryopithecus.* Algunos individuos pertenecientes a esta lejana población de *Sivapithecus* presentan un macizo facial comparable al del gorila; es raro encontrar individuos con una cara de pequeño tamaño, como la del chimpancé.

A finales del Mioceno, hace unos 10 o 12 millones de años, la mayor parte de Eurasia estaba cubierta de bosques.



TRES MANDIBULAS FOSILES, vistas por la norma posterior, para poder observar las prominencias óseas indicadas en los cortes transversales de la derecha. En el corte se observa el plano en que coinciden las mitades derecha e izquierda de la mandíbula. Las prominencias óseas aumentan la resistencia mandibular frente a las tensiones horizontales producidas por la masticación lateralizada. En *Ramapithecus (a)* y *Australopithecus (b)* estas prominencias presentan un desarrollo similar. *Gigantopithecus (c)* recuerda al antropoide del Mioceno llamado *Sivapithecus* y al gorila actual; la eminencia inferior se halla más desarrollada que la superior. En los dos homínidos, el eje mayor de la sección transversal forma un ángulo relativamente marcado con la horizontal; en las mandíbulas que aparecen en la figura, este ángulo es de unos 65 grados, aurque en ocasiones puede ser incluso mayor. Esta característica contrasta con el aspecto que presenta el corte transversal de *Gigantopithecus,* en que la porción anterior es más alargada y se inclina hacia delante, formando un ángulo menor con la horizontal.

Sin embargo, estos bosques no eran de tipo tropical. En consecuencia, la producción de frutos no alcanzaba todo el año, como ocurre en la actualidad en los bosques en que habitan los antropoides. En tales circunstancias, resulta probable que los alimentos disponibles fuesen insuficientes para los antropoides de mayor tamaño. La tendencia a la recolección de alimentos –del tipo de las nueces y raíces en el suelo y siguiendo los límites del bosque– habría podido dar lugar a presiones de selección natural lo suficientes como para favorecer la supervivencia de individuos de mandíbulas más robustas y de dientes dotados de un esmalte más grueso.

Los restos de animales hallados junto a los fósiles de *Ramapithecus* en ciertos lugares de Eurasia indican también una adaptación a los límites boscosos y a la sabana. Así, por ejemplo, los restos de animales hallados en Çandir incluyen especies que habitaban los límites de las zonas boscosas y también las sabanas. En dichas zonas se han encontrado restos de *Ramapithecus*, pero no de otros antropoides. Los restos de animales del Mioceno hallados en Atenas corresponden a especies adaptadas a la sabana y a la estepa. También en este lugar se encontraron restos de *Ramapithecus*, pero no de antropoides. En Hungría, en Pakistán y en la India, los animales y plantas fósiles encontrados junto a los restos de *Ramapithecus* indican que estos homínidos vivían en un ambiente subtropical o por lo menos cálido, junto con ciertos antropoides. En un ambiente de este tipo habrían podido coexistir no sólo *Ramapithecus*, sino también otras especies de mayor tamaño como *Sivapithecus* y *Gigantopithecus*, cuyas proporciones eran parecidas a las de chimpancés y gorilas. Estos primates del Mioceno habrían sufrido un cambio dietético, que fue el que dio origen a su nuevo sistema dental. Durante el período en que vivieron en el borde de las zonas boscosas, pasaron de alimentarse en los árboles a hacerlo en el suelo. De hecho, *Ramapithecus* ocupaba probablemente la sabana y los bosques poco tupidos.

No parece necesario pensar en la posibilidad de que se produjeran dos adaptaciones distintas a las mismas circunstancias, una en *Ramapithecus* y la otra en los grandes dryopithecinos de Eurasia para explicar estos cambios anatómicos. Resulta más simple pensar en la aparición de un proceso de divergencia en un grupo de antropoides adaptados a vivir en la sabana (moverse sobre el suelo y comer alimentos pequeños y duros), y que ocupaban Eurasia hace unos 15 millones de años, la cual dio origen a dos ramas distintas, una que produjo seres de tipo antropoide y la otra que alumbró a individuos de tipo humano. Aunque ambos grupos se hayan extinguido, puede que los *Australopithecus* apareciesen a partir de *Ramapithecus*.



YACIMIENTO DE FOSIL EN LA INDIA, cercano al lugar en que G. Edward Lewis descubrió por vez primera los restos de *Ramapithecus*, en 1932, durante una visita realizada por un grupo de paleontólogos de la Universidad de Yale en 1969. Este yacimiento, situado en las colinas de Siwalik, y abundante en fósiles del Mioceno, contenía dientes del homínido *Ramapithecus*.

Entre ambos existe un período desconocido que abarca unos cuatro millones de años, tiempo muy superior al de la historia del género *Homo*. Aunque, hasta el momento, los fósiles de *Australopithecus* se circunscriben al continente africano, es posible que los dos géneros, *Ramapithecus* y *Australopithecus* hayan sido igualmente cosmopolitas. Sin embargo, esta hipótesis no podrá ser demostrada mientras no se encuentren restos fósiles que la confirmen. A este respecto podemos citar otra situación totalmente superponible. Los primeros fósiles correspondientes a *Homo erectus*, antepasado inmediato del hombre actual, fueron hallados por Eugène Dubois, en Java, en 1890. En 1930 se encontraron fósiles del mismo tipo en China, lo que permitió demostrar que *Homo erectus* se hallaba también distribuido en otras zonas. Mucho más tarde se encontraron fósiles de *Homo erectus* al sur, este y norte de Africa, y también en Europa, lo que pone de manifiesto la amplia distribución de esta especie. En consecuencia, no sería de extrañar que se recogiesen fósiles de *Australopithecus* en otros continentes. Sí resultaría sorprendente que los *Australopithecus* procediesen tan sólo de parte de la población africana de *Ramapithecus*.

Hasta el momento, el estudio de esta vía evolutiva ha permitido pasar de los hominoides a los homínidos, a través de una rama desconocida de *Dryopithecus* que dio origen tanto a los grandes antropoides del sur de Eurasia como a los homínidos de tipo *Ramapithecus*. A su vez, este homínido pudo haber sido el antecesor de los *Australopithecus*. ¿Cuál fue el origen del género *Homo*?

También aquí hay que decir que existe cierto parecido entre *Ramapithecus* y las especies más primitivas del género *Homo*, aunque estas semejanzas no sean siempre las mismas. La denominación *Homo habilis*, creada por Leakey, en 1964, para clasificar ciertos fósiles procedentes del famoso yacimiento de Olduvai Gorge, en Tanzania, ha sido aplicada a un grupo de restos que muchos paleontólogos consideran anteriores a *Homo erectus*. Es posible que deban incluirse en este grupo los fósiles prehumanos hallados por Mary Leakey, en Tanzania, y por Johanson, en Etiopía, cuya antigüedad se calcula en unos tres o cuatro millones de años. Los restos procedentes de Tanzania y de Etiopía corresponderían a miembros primitivos del género *Homo*, y se parecen a otros fósiles descubiertos al nordeste de Kenya por el hijo de Leakey, Ricard E. Leakey, y entre los que se incluye el famoso cráneo, acompañado de su correspondiente macizo facial, y conocido por E. R. (East Rudolf) 1470.

Homo habilis es una especie que merece ser mejor estudiada. Hasta el momento, los especialistas no han llegado a un acuerdo sobre los restos que deben incluirse en esta especie. Sin embargo, parece claro que estos miembros primitivos del género *Homo* poseían caninos e incisivos de mayor longitud que sus contemporáneos, los *Australopithecus*.

Existen también otras diferencias entre los primeros *Homo* y los *Australopithecus*. Tanto el cerebro como el cuerpo de los individuos del género *Homo* eran de mayor tamaño que los correspondientes a los *Australopithecus*. Sin embargo, la posibilidad de que *Homo* proceda de *Australopithecus* ha sido ampliamente discutida, y no vamos a tratarla aquí. Lo que sí discutiremos es el parecido existente entre *Ramapithecus* y las formas más primitivas del género *Homo*, como *Homo habilis* y *Homo erectus*, características que no se hallan en los restos de *Australopithecus*.

En este caso, los datos existentes se basan principalmente en el estudio de mandíbulas y dientes. El parecido citado fue señalado por vez primera por Von Koenigswald al estudiar un resto de *Homo erectus* hallado en el yacimiento de Sangiran, en Java, en 1939. Este fósil ha sido denominado Sangiran II o Pithecanthropus IV. Al igual que ocurre en los restos más primitivos de *Homo* procedentes de Kenya, de Etiopía y de Tanzania (y en el primer resto de *Homo habilis* de Leakey), los caninos superiores de Sangiran II son de tamaño relativamente grande. Además, mientras que los premolares de *Australopithecus* presentan un tamaño parecido al de los molares, en Sangiran II dichas piezas son menores. El canino superior de Sangiran II se superponía a los dientes de la mandíbula inferior, hasta tal punto que dieron lugar a la aparición en los mismos de facetas desgastadas que afectaban tanto al canino como al primer premolar. Junto al canino superior existía un espacio o diastema que lo separaba del incisivo correspondiente.

Estos espacios situados entre los dientes se han considerado siempre como típicos de los antropoides. Durante muchos años, el diastema existente en el fósil Sangiran II había sido interpretado como una anomalía. Sin embargo, uno de los restos recientes más completos procedente de Etiopía, y consistente en un paladar superior íntegro con sus dientes, ha permitido demostrar que este tipo de relación entre los dientes no es anormal en absoluto, ya que la adaptación de los dientes de la mandíbula superior resulta totalmente comparable con la observada en *Ramapithecus*. En cambio, los caninos de *Australopithecus* son tan pequeños que no llegan a superponerse a los dientes de la mandíbula inferior, ni los desgastan, y, por otra parte, tampoco presentan el diastema descrito en la mandíbula superior. Hay que señalar que, en aquellos casos en que los restos de *Homo erectus* y *Homo habilis* (y de *Ramapithecus*) presentan un diastema en la mandíbula superior, los dientes de la mandíbula inferior se encuentran muy apretados. Este tipo de disposición dental es típica de los *Australopithecus*, y no se observa en los antropoides.

Existe también otro dato. Los primeros fósiles del género *Homo* procedentes de Africa, como E. R. 1470, en caso de que realmente correspondan a dicho género, presentan un aspecto craneal y facial que recuerda extraordinariamente a *Australopithecus*. Si no fuera por la presencia de características típicamente humanas, como un mayor tamaño cerebral y una disposición dental parecida a la de *Homo erectus*, estos fósiles podrían clasificarse como correspondientes a una nueva especie de *Australopithecus*, que habría podido ser denominada *Australopithecus habilis*.

El mejor medio para resolver este problema sería partir de la hipótesis de que el grupo de homínidos que dio origen a *Homo* y a *Australopithecus* era más parecido a *Ramapithecus* que otros ejemplares posteriores de *Australopithecus*. La afinidad se basa principalmente en la presencia de caninos de gran tamaño. El único período posible para la aparición de un homínido de este tipo durante el proceso evolutivo correspondería al espacio transcurrido entre la existencia de los últimos ejemplares de *Ramapithecus* y la aparición de los primeros ejemplares de *Homo* y de *Australopithecus*. Este período duró unos cuatro millones de años, y constituye el espacio vacío de mayor duración en la historia evolutiva de la especie humana.

Existe otra posible explicación para resolver el problema, planteada por L. S. B. Leakey. Según esta teoría, *Homo* y *Australopithecus* se habrían separado independientemente a partir de *Ramapithecus*. La tercera posibilidad, mucho menos probable, no concuerda con los datos que hemos planteado en este artículo acerca de *Ramapithecus*. Según dicha interpretación, los antropoides del Mioceno habrían dado origen no sólo a *Ramapithecus*, sino también a otras líneas de homínidos. Si, en el futuro, se encuentran restos que permitan confirmar esta última hipótesis, las ideas actuales sobre el origen de los homínidos deberán ser totalmente revisadas. Aun así, sin embargo, *Ramapithecus* seguirá siendo uno de los ejemplares más primitivos de la línea de los homínidos.

Semiconductores amorfos

Hay materiales vítreos que actúan como conmutadores electrónicos y rivalizan para ciertos usos con los materiales cristalinos. Su comportamiento eléctrico ya se comprende mejor, lo que abre el camino a una amplia serie de nuevos usos

David Adler

Han transcurrido más de 15 años desde la primera demostración de dispositivos de conmutación y memoria basados en los semiconductores amorfos conocidos como vidrios calcógenos, es decir, vidrios que contienen un alto porcentaje de uno o más de los elementos "calcógenos" (azufre, selenio y teluro). Aun cuando la idea de utilizar materiales amorfos como semiconductores en dispositivos electrónicos, en lugar de materiales cristalinos cuidadosamente preparados, resultaba muy prometedora, en un principio el desarrollo de dichos dispositivos progresó muy lentamente debido a la falta de una adecuada base teórica. Sin embargo, durante estos últimos años, se ha disipado gran parte del misterio que rodeaba el comportamiento de los electrones en sólidos amorfos. Posiblemente, el concepto más importante surgido al respecto consiste en que las propiedades electrónicas de los semiconductores amorfos están controladas por "centros activos" bien definidos, entendiendo por tales a ciertas regiones ocupadas por átomos en estados energéticos peculiares, tal como ocurre con respecto a las propiedades electrónicas de los semiconductores cristalinos. Durante el mismo período se han producido avances importantes en la comprensión del mecanismo de conmutación en vidrios calcógenos, y también en el desarrollo de dispositivos comerciales basados en dicho mecanismo. El rápido progreso alcanzado en un tiempo bastante breve promete reavivar el interés hacia un campo de trabajo del que a veces se dice que es la última frontera de la física del estado sólido.

La principal característica diferenciadora de un cristal es que sus átomos o moléculas constituyentes están alineados en ordenamientos tridimensionales con una periodicidad de largo alcance. Debido precisamente a las simplificaciones matemáticas que se derivan de dicha periodicidad, se llegó a entender perfectamente los fenómenos físicos que ocurren en sólidos cristalinos, poco después de desarrollarse, hace 50 años, la mecánica cuántica. Y este conocimiento detallado condujo con el tiempo a la creación de diferentes dispositivos de estado sólido tales como transistores, osciladores de microondas y láseres. A pesar de estos grandes triunfos, las delicadas técnicas de preparación y crecimiento de cristales, necesarias para la fabricación de los diferentes dispositivos basados en materiales cristalinos, suponen importantes inconvenientes económicos. Un buen ejemplo de ello lo tenemos en la preparación de células solares, dispositivos que convierten directamente la radiación solar en energía eléctrica. La fiabilidad de la tecnología de las células solares ha sido ampliamente demostrada a través de su utilización en vehículos espaciales. El material más importante que se necesita para la fabricación de células solares convencionales es el silicio, que ocupa el segundo lugar entre los elementos más abundantes en la corteza terrestre. Sin embargo, es improbable que en un futuro inmediato se llegue al uso en gran escala de tales células solares, a pesar de la urgente necesidad de nuevos recursos energéticos. El mayor obstáculo en este sentido estriba en la preparación de silicio en forma cristalina de alta calidad. Usando la tecnología actual, el coste de producción de electricidad con células solares es del orden de 1,50 dólares por kilovatio-hora, 100 veces mayor que el de la electricidad a partir de combustibles fósiles.

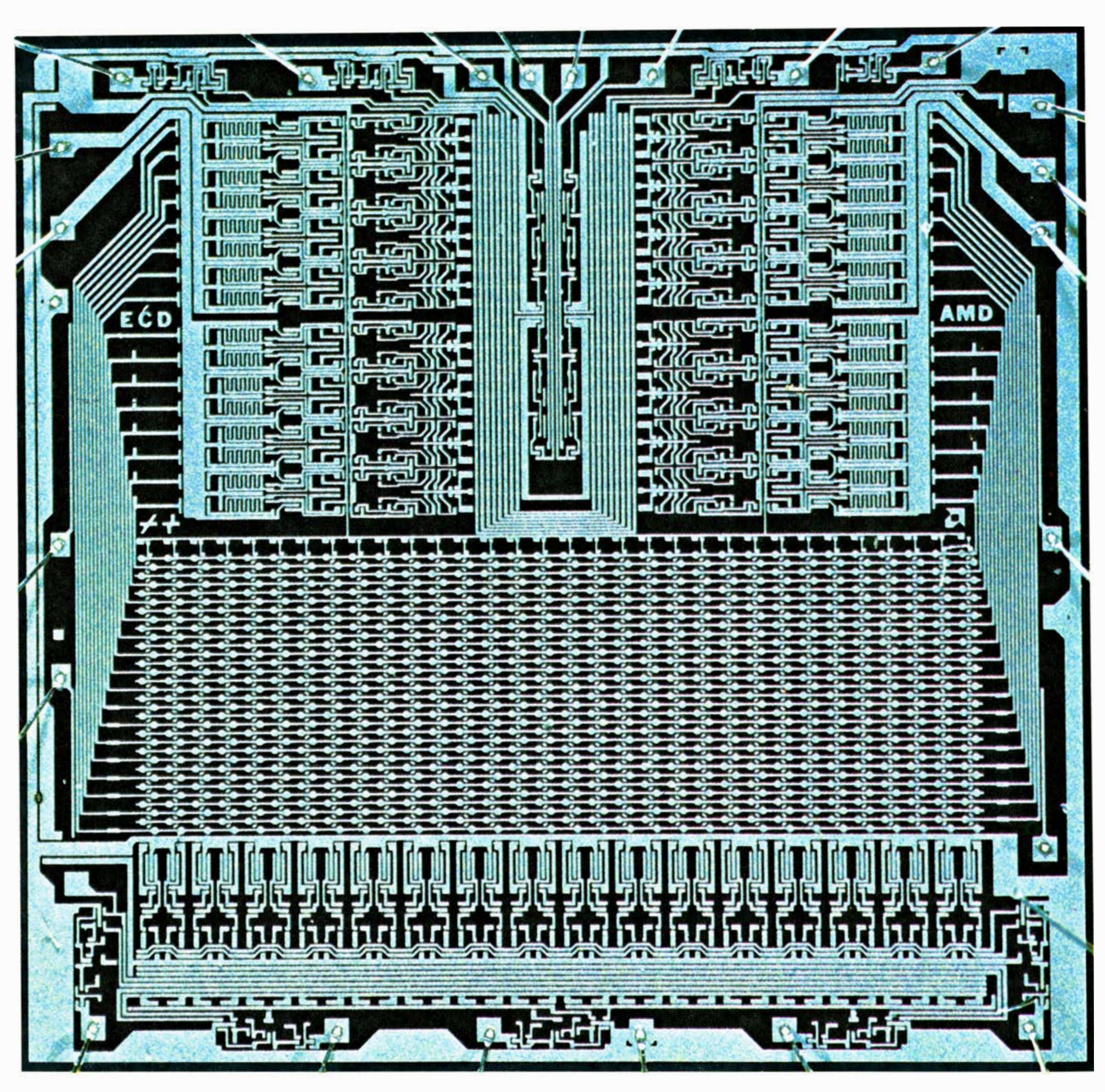
Desde la perspectiva general de la tecnología industrial, los sólidos cristalinos altamente periódicos ocupan un enclave muy pequeño, aunque también muy importante. Los sólidos no periódicos, o amorfos, constituyen una clase de materiales con una mayor diversidad de propiedades físicas. Por otra parte, su preparación no requiere en general el empleo de técnicas de crecimiento cuidadosamente controladas, lo cual supone una ventaja económica vital para muchas aplicaciones. Los sólidos amorfos suelen prepararse fácilmente en forma de vidrios, enfriando rápidamente un material a partir de su fase líquida. A una diversidad aun mayor de composiciones se puede dar la forma de películas delgadas permitiendo que su vapor se deposite sobre un sustrato frío.

La complejidad matemática del tratamiento de un sistema no periódico impidió la comprensión teórica de las propiedades microscópicas de los semiconductores amorfos, retrasando seriamente su posible uso como dispositivos electrónicos. Una importante excepción la constituyó el campo de la electrofotografía, en el que la Xerox Corporation aprovechó las propiedades fotoconductoras de ciertos vidrios con una base de selenio para poner en pie una industria multimillonaria. Pero incluso este notable éxito sólo se produjo más de 20 años después de que Chester F. Carlson describiera el concepto fundamental en 1938. El proceso no ha sido entendido del todo ni siquiera en la actualidad. La falta de comprensión se convirtió en un obstáculo todavía más serio para Stanford R. Ovshinsky, que inició su trabajo como pionero en el campo de los semiconductores amorfos en el año 1958. Sus trabajos, publicados en 1968, en los que describe su descubrimiento anterior de dos tipos de fenómenos de conmutación reversible, suscitó un gran interés, pero el interés decayó considerablemente cuando se comprobó cuán difícil era hallar explicaciones adecuadas para los fenómenos. Voy a describir a continuación las nuevas obser-

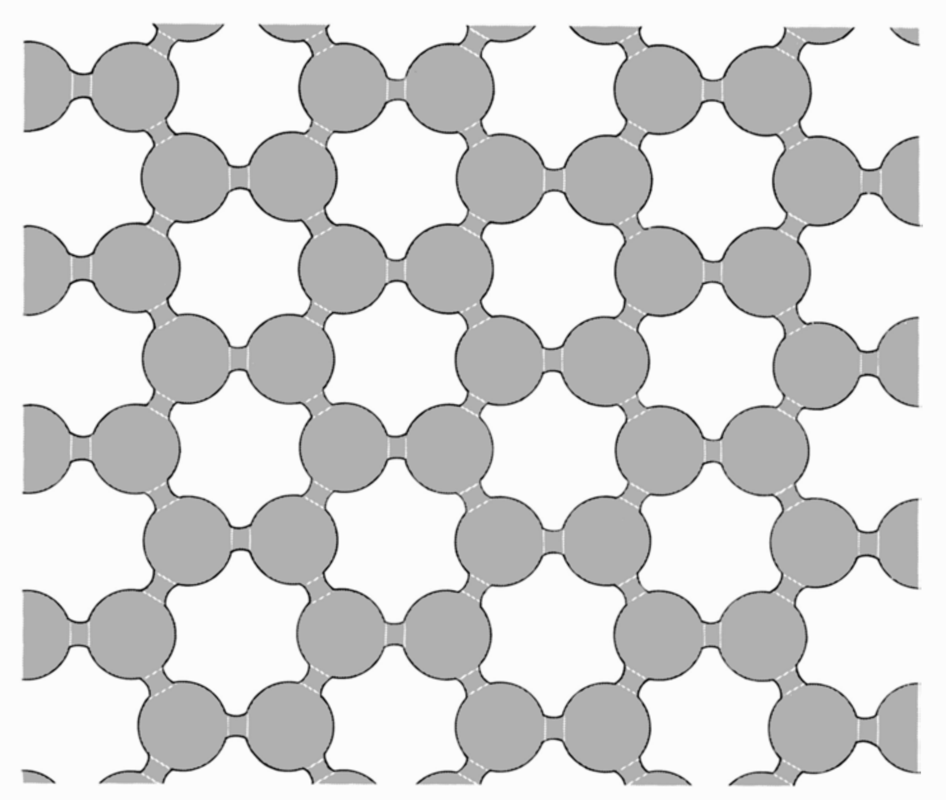
vaciones que han convencido a muchos investigadores en este campo de trabajo de que el largo período de perplejidad ha llegado a su fin.

Tal como sucede con los sólidos cristalinos, los sólidos amorfos pueden ser aislantes, semiconductores o metales, y, en algunos casos, a muy bajas temperaturas, pueden ser incluso superconductores. La teoría cuántica de los sólidos, tal como se elaborara originalmente (y todavía hoy se presenta en los libros de texto), se basa por entero en la existencia de un orden de largo alcance. El comportamiento de los semiconductores está íntimamente relacionado con esta teoría. En consecuencia, durante mucho tiempo se creyó que los sólidos amorfos no podían ser semiconductores. Sin embargo, en 1955, B. T. Kolomiets y sus colaboradores, de la URSS, descubrieron que los vidrios calcógenos podían actuar como semiconductores. En un esfuerzo por explicar este inesperado comportamiento, los científicos rusos A. F. Ioffe y A. R. Regel sugirieron, en 1960, que las propiedades electrónicas de los vidrios calcógenos tienen su origen en un orden de corto alcance y no en un orden de largo alcance, pero a pesar de que muchos abordaron este problema, no se llegó con facilidad a una interpretación detallada del mismo.

En la actualidad, muchos años después de ese descubrimiento, resulta ya evidente que la clave de la comprensión de las propiedades de los sólidos no es su estructura periódica, sino la naturaleza química de sus átomos constituyentes. La periodicidad de los sólidos cristalinos simplifica enormemente los cálculos cuantitativos, pero no es esencial para la explicación de las características principales de su comportamiento eléctrico, óptico y magnético. Un plantea-



MEMORIA "OVONIC", que tiene una capacidad de almacenamiento de 1024 bits y mide 5,6 milímetros de lado. Los elementos de memoria de semiconductor amorfo constituyen la figura regular situada en la mitad inferior de la célula. Los excitadores y decodificadores, que utilizan la tecnología convencional del cristal de silicio, ocupan la otra mitad. Este dispositivo se utiliza como una memoria alterable en ordenadores.



SOLIDO CRISTALINO, representado esquemáticamente en dos dimensiones. En este hipotético material, cada átomo forma tres enlaces idénticos con sus vecinos más próximos; en consecuencia, los enlaces de cada átomo forman entre sí ángulos de 120 grados. Se puede decir que la estructura periódica resultante, consistente en hexágonos regulares, muestra un orden de largo alcance.

miento más fructífero consiste en intentar comprender la estructura electrónica de los átomos situados en columnas diferentes del sistema periódico de los elementos, antes de pretender analizar el problema, mucho mayor, de sólidos que contienen del orden del 10^{24} átomos.

Como paso inicial en esta dirección vamos a limitarnos a aquellos elementos que no son ni tierras raras ni metales de transición, es decir, los elementos que aparecen en las columnas I a VIII de la tabla periódica ordinaria *(véase la figura de la página 26)*. Un resultado fundamental de la teoría cuántica es que sólo es posible un número discreto de estados para los electrones de cualquier átomo o molécula, y que estos estados se pueden agrupar en capas, en las cuales la energía de los electrones presenta valores comprendidos dentro de un pequeño intervalo. La diversidad de elementos en nuestro universo es una consecuencia del principio de exclusión, enunciado por vez primera por Wolfgang Pauli, que prohíbe que dos electrones cualesquiera de un mismo átomo ocupen un mismo estado. Las propiedades de cada átomo están determinadas fundamentalmente por el número de electrones que posean. El número de estados disponibles en cada capa adquiere una importancia crucial. Para los átomos objeto de la presente discusión, la capa más externa (es decir, la capa de energía más elevada) tiene siempre ocho estados posibles. Los ocho estados de esta capa pueden subdividirse ulteriormente en dos grupos: dos estados de energía más baja, o estados *s*, y seis estados de energía más alta, o estados *p*. De hecho, la separación energética entre los estados *s* y *p* es bastante pequeña. Las propiedades de cada átomo vienen determinadas muy especialmente por el número de electrones existentes en la capa más externa; número que explica las ocho columnas de la tabla periódica.

Cada uno de los ocho estados posibles posee una naturaleza característica que difiere de todas las demás. Dichos estados pueden identificarse por su momento cinético orbital (análogo al referente a la trayectoria de un planeta en su movimiento de traslación alrededor del sol) y su momento cinético intrínseco o espín (análogo al que se refiere a la rotación del planeta alrededor de su eje). Los estados *s* tienen solamente un valor posible del momento cinético orbital; los estados *p* tienen tres. Sin embargo, todos los electrones tienen sólo dos valores del espín, designados generalmente como espín "hacia arriba" y espín "hacia abajo". Las propiedades magnéticas de los sólidos están determinadas por los espines de los electrones y, por su parte, éstos están relacionados con la naturaleza química del material.

Las combinaciones de átomos constituirán moléculas estables siempre que la energía de la combinación molecular sea menor que la energía de los átomos no combinados. Dicha disminución de energía puede producirse de dos maneras. En aquellas moléculas cuya estabilidad se debe a un tipo de enlaces conocidos como enlaces iónicos, la reducción en la energía surge de la transferencia de uno o más electrones de un átomo a otro y la subsiguiente atracción electrostática entre los átomos con cargas opuestas que aparecen como consecuencia de dicha transferencia de electrones. En las moléculas cuya estabilidad se debe a los denominados enlaces covalentes, la disminución de energía tiene un origen más sutil: los electrones exteriores de los átomos adyacentes se reparten entre los dos átomos. Siempre que dos átomos cualesquiera se encuentren lo suficientemente próximos como para que sus electrones exteriores interaccionen, las energías de sus estados pueden resultar alteradas por la interacción. Generalmente, algunos de los estados reducen su energía, en tanto que otros, en número igual a los anteriores, la aumentan; los primeros son estados "enlazantes" y los segundos son "antienlazantes". Cuando los estados enlazantes están llenos (siempre por pares de electrones cuyos espines son opuestos) y los estados antienlazantes están vacíos, la molécula resulta ligada en su grado máximo. Aquellos estados cuya energía no sufra una alteración apreciable se conocen como estados "no enlazantes". Su ocupación no altera la estabilidad de la molécula.

Los estados *s* son no direccionales; los estados *p* poseen una direccionalidad definida (generalmente los tres estados *p* son perpendiculares entre sí). Debido a esta direccionalidad, los enlaces entre electrones en estados *p* suelen ser más fuertes que los correspondientes a electrones en estados *s*. Un resultado todavía más importante es el de que un par de electrones (uno de ellos con espín hacia arriba y el otro con espín hacia abajo) no pueda disminuir su energía al enlazarse covalentemente con otro par. Esto parecería implicar, por ejemplo, que los elementos de la columna II del sistema periódico no pueden formar enlaces covalentes. En realidad, dado que la separación energética entre los estados *s* y *p* es generalmente del mismo orden de magnitud que la energía de enlace, es posible con frecuencia "promocionar"

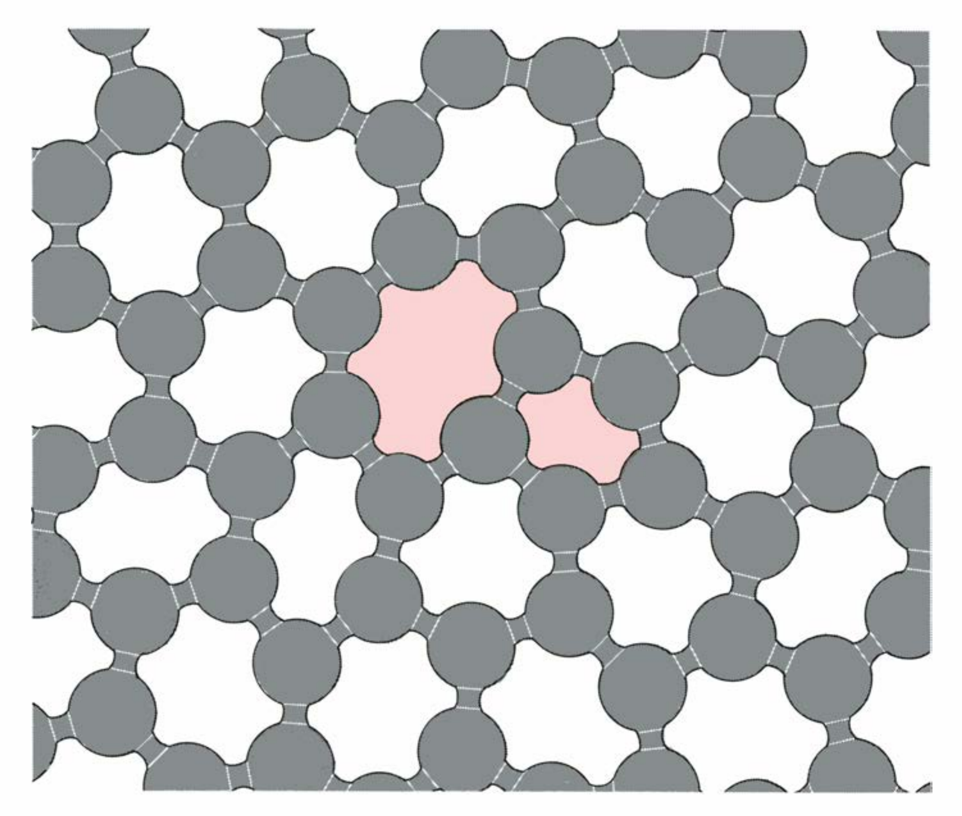
uno de los electrones *s* apareados a un estado *p*, con lo cual se forman dos enlaces covalentes en lugar de ninguno. Esto se puede llevar a cabo hasta para los elementos de la columna IV, en la cual la configuración sp^3 da lugar al máximo número posible de enlaces *s-p*, es decir, cuatro.

Al iniciar la columna V (elementos pnictidos), ya no resulta energéticamente favorable promocionar uno de los electrones *s*, debido a que ya hay tres electrones *p* desapareados sin la promoción y, por tanto, no se pueden formar enlaces adicionales. Los electrones *s* permanecen apareados (estos electrones constituyen lo que se suele denominar un par aislado) y sólo son posibles tres enlaces *p*. Los elementos de la columna VI (calcógenos) poseen generalmente un par *p* aislado además del par *s* aislado, y, por consiguiente, la configuración de menor energía incluye solamente dos enlaces *p*. Análogamente, los átomos de la columna VII (halógenos) tienen dos pares *p*, dejando únicamente un electrón *p* disponible para enlace. Finalmente, los elementos de la columna VIII son esencialmente inertes y no forman enlaces covalentes en condiciones normales. Además de lo antedicho, la direccionalidad de los enlaces *p* conduce a valores fijos de los ángulos de enlace en todos los elementos situados en las columnas II a VI.

Un sólido puede considerarse como una molécula gigante constituida en el caso típico por unos 10^{24} átomos. Al igual que las moléculas, los sólidos pueden existir únicamente si la energía de todo el sistema disminuye en el curso de su formación. El origen de esta disminución de la energía es totalmente análogo al origen de la misma en las moléculas simples. Sin embargo, el sistema es mucho más difícil de describir mediante ecuaciones, puesto que es necesario tomar en consideración unos 10^{24} estados, en vez de manipular sólo unos cuantos. En cualquier caso, es evidente que los estados discretos de los átomos aislados quedan desplazados por las interacciones interatómicas de los electrones. Debido al enorme número de estados existentes en un pequeño intervalo de energías, es posible tratarlos como si sus energías estuvieran distribuidas continuamente dentro de dicho intervalo. En lugar de caracterizar cada uno de los estados, como sucede en un átomo o en una pequeña molécula, se suele trazar en un gráfico el número de estados por unidad de volumen y por unidad de energía (esta magnitud se denomina densidad de estados), en función de la energía. Puesto que todos los estados tienen su origen en niveles atómicos discretos, en el caso de cualquier sólido existen siempre ciertas regiones de energías en las cuales existen dichos estados (bandas), separadas por regiones en las que no es posible su existencia (huecos o "gaps"). Todo sólido contiene un número fijo de electrones, *N*, que tienen que ocupar los estados permitidos. Debido al principio de exclusión, la energía más baja de cualquier sólido se alcanza cuando los electrones llenan los *N* estados más bajos, dejando vacíos todos los estados superiores. Convencionalmente, en el diagrama se suele trazar una línea llamada nivel de Fermi (en honor de Enrico Fermi), situada a mitad de camino entre el estado lleno más elevado y el estado vacío más bajo.

Se puede demostrar fácilmente que en un sólido cristalino la periodicidad de largo alcance tiene dos importantes consecuencias: en primer lugar, existen ciertas energías, denominadas extremos o límites de bandas, en las cuales la densidad de estados electrónicos disminuye bruscamente hasta anularse; en segundo lugar, todos los estados electrónicos de las bandas se extienden a lo largo de todo el material. Esta última propiedad es enormemente importante, puesto que de ella se derivan directamente las elevadas conductividades eléctricas que son características de los metales. La teoría usual de los sólidos se basa enteramente en la suposición de que la periodicidad de largo alcance es perfecta o casi perfecta. La mera existencia de sólidos amorfos ha supuesto siempre una grave dificultad para esta interpretación, pero el descubrimiento de que estos cuerpos pueden mostrar el mismo tipo de propiedades eléctricas que los sólidos cristalinos ha conducido finalmente a una reconsideración crítica de la importancia real de la periodicidad. En este sentido, el estudio de los materiales amorfos continúa teniendo una repercusión decisiva sobre nuestra comprensión de los materiales cristalinos.

La conducción eléctrica exige la presencia de partículas cargadas que puedan moverse con facilidad en presencia de un campo eléctrico externo. La conductividad misma es proporcional al producto del número de tales portadores existentes por unidad de volumen (o concentración) por la velocidad media que adquieren por unidad del campo aplicado,



SOLIDO AMORFO del mismo material hipotético que hemos representado en la figura precedente; estaría constituido predominantemente por hexágonos irregulares. Los requisitos del enlace trivalente de cada átomo quedan satisfechos localmente y los enlaces tienen la misma longitud que poseían en el cristal. Sin embargo, los ángulos pueden variar hasta 20 grados, en más o en menos, respecto al valor medio de 120 grados. Por otra parte, diseminados entre los hexágonos, existen algunos anillos con cinco y siete miembros e incluso algunos con ocho miembros. Aunque en los sólidos amorfos se observa la presencia de un orden de corto alcance (puesto que se conoce la distancia entre los átomos), puede advertirse la ausencia de un orden de largo alcance.

COLUMNA	I	II	III	IV	V	VI	VII	VIII
NOMBRE DE LA FAMILIA	METALES ALCALINOS	ALCALINO-TERREOS			PNICTIDOS	CALCOGENOS	HALOGENOS	GASES NOBLES
ELEMENTOS	LITIO, SODIO, POTASIO, RUBIDIO, CESIO	BERILIO, MAGNESIO, CALCIO, ESTRONCIO, BARIO	BORO, ALUMINIO, GALIO, INDIO TALIO	CARBONO, SILICIO, GERMANIO, ESTAÑO, PLOMO	NITROGENO, FOSFORO, ARSENICO, ANTIMONIO, BISMUTO	OXIGENO, AZUFRE, SELENIO, TELURO, POLONIO	FLUOR, CLORO, BROMO, IODO, ASTATO	NEON, ARGON, KRIPTON, XENON, RADON
NUMERO DE ELECTRONES EN LA CAPA MAS EXTERNA	1	2	3	4	5	6	7	8
CONFIGURACION ELECTRONICA DE ENERGIA MAS BAJA	s ↑	s^2 ↑↓	s^2 p ↑↓ ↑	s^2 p^2 ↑↓ ↑↑	s^2 p^2 ↑↓ ↑↑↑	s^2 p^4 ↑↓ ↑↑↑↓	s^2 p^5 ↑↓ ↑↑↑↓↓	s^2 p^6 ↑↓ ↑↑↑↓↓↓
CONFIGURACION ELECTRONICA PARA ENLACE MAXIMO	s ↑	sp ↑↑	sp^2 ↑↑↑	sp^3 ↑↑↑↑	p^3 ↑↑↑	p^4 ↑↑↑↓	p^5 ↑↑↑↓↓	p^6 ↑↑↑↓↓↓
NUMERO MAXIMO DE ENLACES POR ATOMO	1	2	3	4	3	2	1	0
ANGULO DE ENLACE OPTIMO	—	180°	120°	109.5°	90°-100°	100°-105°	—	—
CONFIGURACION GEOMETRICA	MOLECULA DIATOMICA	CADENA	LAMINA	ESTRUCTURA RIGIDA	LAMINA ARRUGADA (O CINTA)	CADENA ARRUGADA (O ANILLO)	MOLECULA DIATOMICA	—
DIMENSIONALIDAD DEL SOLIDO LIGADO	CERO	UNA DIMENSION	DOS DIMENSIONES	TRES DIMENSIONES	DOS DIMENSIONES	UNA DIMENSION	CERO	CERO
ENERGIA DE COHESION DEL SOLIDO COVALENTE	MUY PEQUEÑA	PEQUEÑA	MODERADA	MUY GRANDE	GRANDE	MODERADA	PEQUEÑA	NULA

SISTEMA PERIODICO DE LOS ELEMENTOS, en forma abreviada, que muestra cómo los electrones añadidos a las capas más exteriores de los átomos en las sucesivas columnas de la tabla están divididos entre los ocho estados posibles: dos estados *s* de energía más baja y seis estados *p* de energía más alta. (La tabla omite todos los metales de "transición", tales como hierro, cobre, plata y oro, así como todos los elementos de las tierras raras.) Los estados *s* tienen un solo valor posible del momento angular orbital; los estados *p* tienen tres. De aquí se desprende que los estados *s* estarán llenos cuando contengan un electrón con espín "hacia arriba" y otro con espín "hacia abajo", tal como indican las flechas. Los tres estados *p* pueden acomodar un total de tres electrones con espín hacia arriba y otros tres con espín hacia abajo. Las asignaciones de los espines de los electrones aparecen en la segunda fila. Las configuraciones correspondientes al máximo enlace covalente entre los átomos se dan en la tercera fila. Para los elementos de las columnas II, III y IV, uno de los electrones del estado *s* se promociona a un estado *p*, formando estados híbridos *(color)*, cada uno de los cuales puede formar un enlace. Así, el carbono, en la columna IV, puede formar cuatro enlaces, correspondientes a los cuatro estados desapareados sp^3. Los elementos de la columna V tienen un máximo de tres enlaces, todos estados *p*. En la columna VI el apareamiento de dos electrones en el estado *p* deja sólo dos estados *p* para enlaces. En la columna VII solamente está desapareado un electrón en el estado *p*, y en la VIII ninguno. Los elementos de la IV forman estructuras rígidas tridimensionales. Cuando dos o más elementos constituyen combinaciones amorfas, cada átomo mantiene localmente sus propiedades de enlace características.

una magnitud denominada movilidad del portador. Para lograr la conducción metálica, se precisa una gran movilidad y una gran concentración de portadores. Cuando una banda de energías está parcialmente llena, los electrones presentes en ella pueden contribuir a la conducción redistribuyéndose adecuadamente entre los estados vacíos de la banda; como consecuencia, los sólidos cristalinos con bandas parcialmente llenas son metálicos. Por otro lado, los electrones de una banda llena no pueden redistribuirse de forma significativa debido al principio de exclusión y, por tanto, en tal caso no se produce conducción. De este modo, los sólidos son metales o aisladores según que el nivel de Fermi se sitúe respectivamente en el centro de una banda o en un hueco.

Los semiconductores son simples aisladores en los que la anchura de dicho hueco es relativamente pequeña: menor que unos 3 electrón-voltios. En tales casos, la energía térmica es suficiente para promocionar algunos de los electrones situados en la banda llena más alta (banda de valencia) a la banda vacía más baja (banda de conducción), en la que estos electrones se pueden mover libremente al aplicarse un campo externo. Ahora bien, cuando algunos electrones son llevados a la banda de conducción, dejan tras de sí un número igual de estados vacíos o "agujeros" cerca de la cumbre de la banda de valencia. Estos estados vacíos actúan como si fueran portadores cargados positivamente. Tanto los electrones como los agujeros pueden contribuir a la conducción en un semiconductor y a temperaturas elevadas ambos pueden ser excitados a partir de la energía térmica existente en el interior del material.

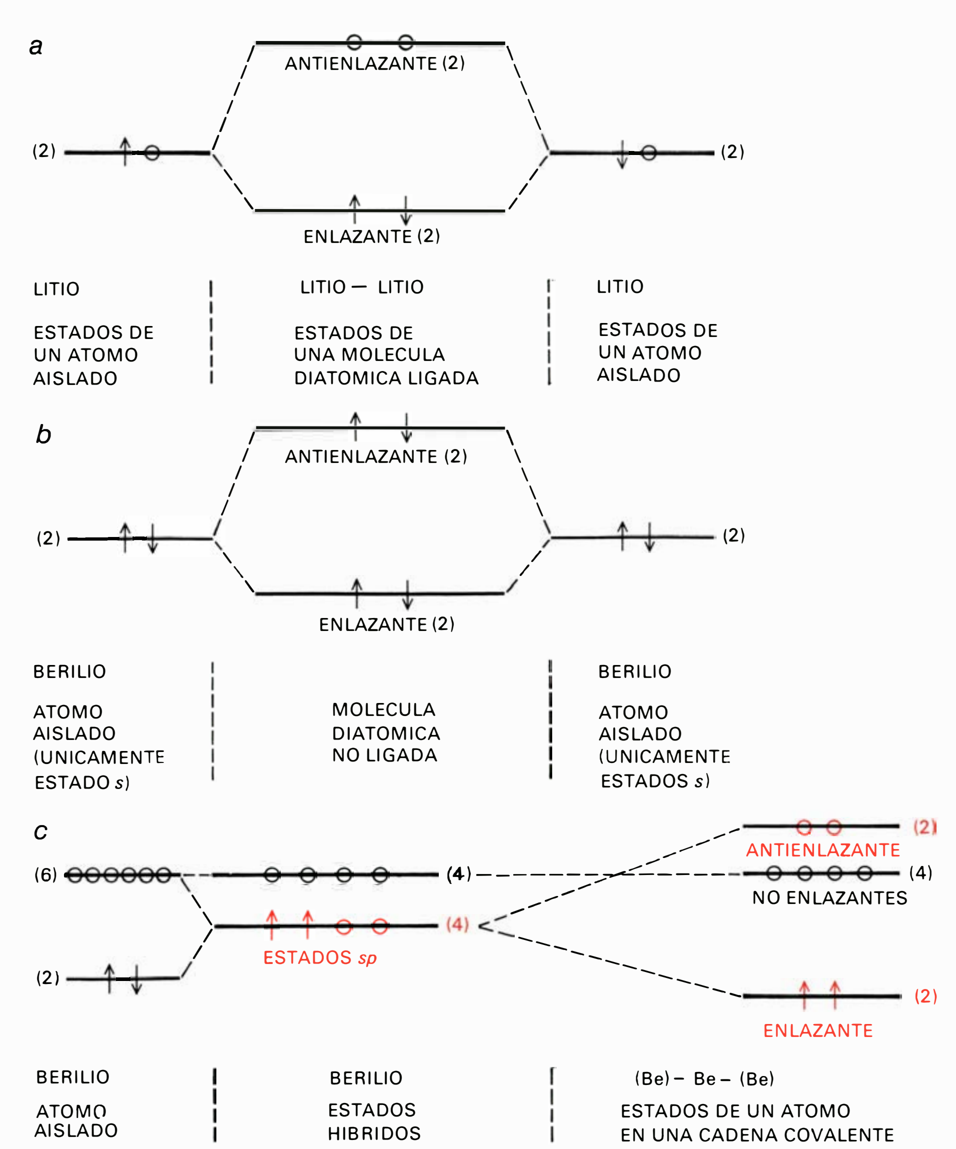
Las aplicaciones electrónicas más importantes de los semiconductores cristalinos no dependen de los portadores excitados térmicamente, sino de los portadores que existen debido a la presencia de átomos de impurezas en el sólido. Un caso típico es el del silicio al que se ha agregado una pequeña cantidad de fósforo. El fósforo tiene exactamente un electrón más que el silicio. Este hecho es muy importante desde el punto de vista químico, puesto que el estado de menor energía de un átomo de silicio, en un material covalente, es aquel en que está rodeado por un tetraedro formado por sus cuatro vecinos más próximos, cada uno de los cuales está ligado covalentemente a uno de los cuatro electrones externos del átomo central. Se obtiene así una estructura electrónica en la cual los ocho estados correspondientes a la capa más externa se separan en niveles enlazantes y antienlazantes *(véase "a" en la figura de la página 29)*. Debido a que el fósforo posee un quinto electrón externo, su estructura enlazante energéticamente preferida es trigonal (con tres vecinos próximos) y no tetraédrica. Sin embargo, cuando se incorporan unos pocos átomos de fósforo a un cristal de silicio, la estructura geométrica del cristal los fuerza a ocupar las mismas posiciones de los átomos de silicio, con lo cual los átomos de fósforo están obligados a enlazarse también tetraédricamente. La banda de conducción del silicio proviene de los estados antienlazantes que generalmente están vacíos. El electrón adicional del átomo de fósforo permanece normalmente en las proximi-

dades, pero su energía es parecida a la de un electrón situado al fondo de la banda de conducción del silicio *(véase "c" en la figura de la página 29)*. El estado resultante presenta, por tanto, dos características principales: en primer lugar, está localizado en el espacio, y, en segundo lugar, puede ser ionizado absorbiendo únicamente una pequeña cantidad de energía, con lo cual se crea un portador "libre" en la banda de conducción. La elaboración de semiconductores cristalinos con una amplia variedad de propiedades eléctricas tiene su origen precisamente en el hecho de que es posible incorporarles diferentes tipos y concentraciones de átomos de impurezas, procedimiento al que se conoce con el nombre de "adulteración selectiva" *(doping)*.

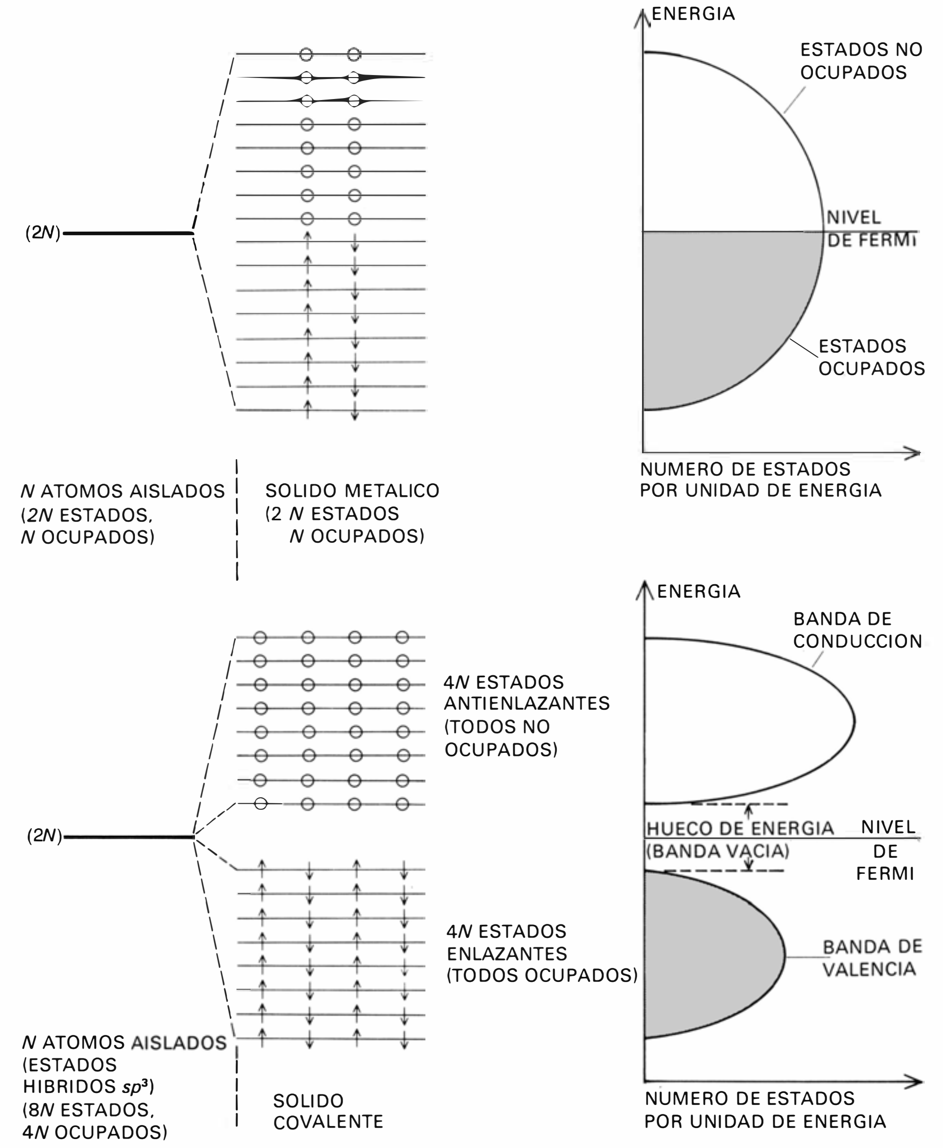
El primer paso semicuantitativo en la generalización de la teoría de los semiconductores cristalinos a los materiales amorfos fue dado por Sir Nevill Mott, de la Universidad de Cambridge. Este investigador observó que los límites precisos de las bandas de los sólidos cristalinos son simplemente una consecuencia de la periodicidad de largo alcance y, por tanto, no deberían existir en los materiales amorfos, siendo reemplazados en éstos por bandas con "colas".

Cabría esperar que estas colas de las bandas llevaran a conducción metálica, pero Mott postuló que los estados situados en ellas no se prolongan a lo largo de todo el sólido, como sucede en los materiales cristalinos puros; por el contrario, estos estados se presentan localizados en el espacio, exactamente igual a como sucede con el estado ocupado por el quinto electrón del fósforo en el silicio cargado con fósforo. Mott sugirió además que, incluso para un sólido amorfo, dichos estados electrónicos se extienden únicamente cuando su densidad supera un determinado valor. Este hecho conduce a la existencia de energías críticas en cada banda, para las cuales se produce un salto acusado en la oportuna movilidad, desde valores despreciables a valores finitos. Estas energías críticas, o "límites de movilidad", desempeñan en los sólidos amorfos el mismo papel que los límites de las bandas en los sólidos cristalinos. La diferencia energética entre los límites de movilidad de la banda de valencia y los de la banda de conducción constituyen el "hueco de movilidad". Al postular la existencia de un hueco de movilidad en sustitución del hueco de bandas de los materiales cristalinos, el modelo de Mott permite explicar claramente cómo es posible que ciertos materiales amorfos se comporten como semiconductores.

El modelo de Mott fue ampliado por Morrel H. Cohen y Hellmut Fritzsche, de la Universidad de Chicago, trabajando en colaboración con Ovshinsky. Su modelo se basa en cuatro principios: primero, los materiales amorfos poseen colas de bandas cuya naturaleza exacta depende de la extensión de las desviaciones de la periodicidad perfecta; segundo, existen límites de movilidad muy bien definidos, que separan los estados localizados y los extendidos en cada banda; tercero, las colas de banda localizadas se solapan en el hueco de movilidad, lo cual da lugar a una densidad de estados finita en el nivel de Fermi; cuarto, puesto que los sólidos amorfos presentan las estrictas limitaciones propias de las estructuras cristalinas, cabe esperar que cada átomo llene sus propias exigencias de valencia localmente, eliminando de esta forma cualquier estructura



ESTADOS ELECTRONICOS QUE FAVORECEN EL ENLACE (*a y c*) **comparados con una configuración electrónica en la que el enlace no se produce** (*b*). **Puesto que los átomos de litio tienen solamente un electrón en su capa externa, los electrones de dos átomos bastan para llenar el estado enlazante, dando lugar a una molécula de litio diatómica** (*a*). **La energía de la molécula es mucho más baja que la de los átomos aislados. (En estos diagramas, las líneas horizontales se refieren a las energías de los diferentes estados.) Los espines de los electrones de la molécula de litio están apareados y, en consecuencia, no se observan efectos magnéticos. El mecanismo responsable del enlace en la molécula de litio no está presente en el átomo de berilio** (*b*), **que tiene dos electrones en su capa externa. Cuando dos átomos de berilio están lo suficientemente cerca como para interaccionar, dos estados** *s* **reducen su energía y dos la aumentan, formando estados "antienlazantes". (En la figura sólo se representan los estados** *s*.) **La energía de los estados antienlazantes aumenta siempre más de lo que disminuye la de los estados enlazantes, de forma que la configuración electrónica no conduce a la disminución general de energía necesaria para formar una molécula diatómica. Sin embargo, los átomos de berilio pueden formar cadenas por el mecanismo de hibridación** *sp* (*c*). **Para promocionar uno de los electrones** *s* **a un estado** *p* **tan sólo se necesita una pequeña cantidad de energía. Puesto que los dos electrones externos ya no necesitan estar apareados, ambos pueden formar enlaces covalentes con sus átomos vecinos más cercanos. Si las dos energías enlazantes son mayores que la simple energía de promoción, se formará una cadena covalente fuertemente ligada. El berilio puede polimerizarse, si bien forma un sólido metálico con energía más baja. Algunos elementos calcógenos (azufre, selenio y teluro) sí se polimerizan.**



LA CONDUCTIVIDAD DE UN SOLIDO depende de la configuración de las bandas de energía que se producen cuando la interacción entre los electrones de átomos próximos da lugar a la diseminación de los estados energéticos de los átomos en una serie de niveles poco separados entre sí. Un sólido metálico tal como el litio *(arriba)* forma una única banda continua de energía en la cual sólo están llenos los estados de la mitad inferior. El nivel de Fermi separa los estados ocupados de los desocupados. Cuando el nivel de Fermi se sitúa dentro de una banda, el sólido es un metal y un conductor. En un sólido enlazado covalentemente, tal como el silicio *(abajo)*, los estados híbridos sp^3 de los átomos aislados se desdoblan en estados enlazantes y antienlazantes que, a su vez, se distribuyen en bandas. La banda enlazante está completamente llena y la banda antienlazante está completamente vacía. El nivel de Fermi se sitúa en el hueco de energía *(gap)*, una región donde no existen estados electrónicos. En tales casos, el sólido es un aislante. Si el hueco de energía es lo suficientemente pequeño como para que la energía térmica pueda llevar a algunos electrones de la banda de valencia a la banda de conducción, la conductividad eléctrica del sólido aumentará con la temperatura. El sólido es entonces un semiconductor. En ambos diagramas, los estados sombreados están llenos. Los estados no sombreados están vacíos en la configuración de energía más baja.

acusada en la densidad de estados localizados en el citado hueco.

Durante los años inmediatamente siguientes, los principales progresos en este campo fueron sobre todo de tipo experimental y no teórico. La imagen resultante es la de que existen tres clases de semiconductores amorfos enlazados covalentemente, con propiedades bastante diferentes entre sí. En un extremo del espectro tenemos a los sólidos amorfos enlazados tetraédricamente, tales como el silicio y el germanio, cuyas propiedades no son radicalmente diferentes de las que presentan los correspondientes cristales. A estos materiales se les pueden añadir ciertas impurezas químicas. Por otra parte, algunos de los espines de los electrones que poseen están desapareados (este hecho da lugar a ciertos efectos magnéticos). Con ellos se pueden preparar diodos de estado sólido.

En agudo contraste con estos materiales están los vidrios calcógenos, que son básicamente insensibles a las impurezas, no poseen en absoluto espines desapareados, excepto bajo condiciones de desequilibrio muy extremas, y, en general, presentan características simétricas de corriente-voltaje. Por otro lado, se puede conseguir que estos vidrios muestren efectos de conmutación y memoria reversibles bajo la acción de campos eléctricos intensos, así como ciertos cambios estructurales al ser iluminados. Los vidrios pnictidos (vidrio cuyo componente básico es un elemento de la columna V del sistema periódico) presentan propiedades intermedias a las de estos dos grupos. Hasta hace poco no existía una interpretación teórica satisfactoria de las diferencias existentes entre las tres clases de semiconductores citadas.

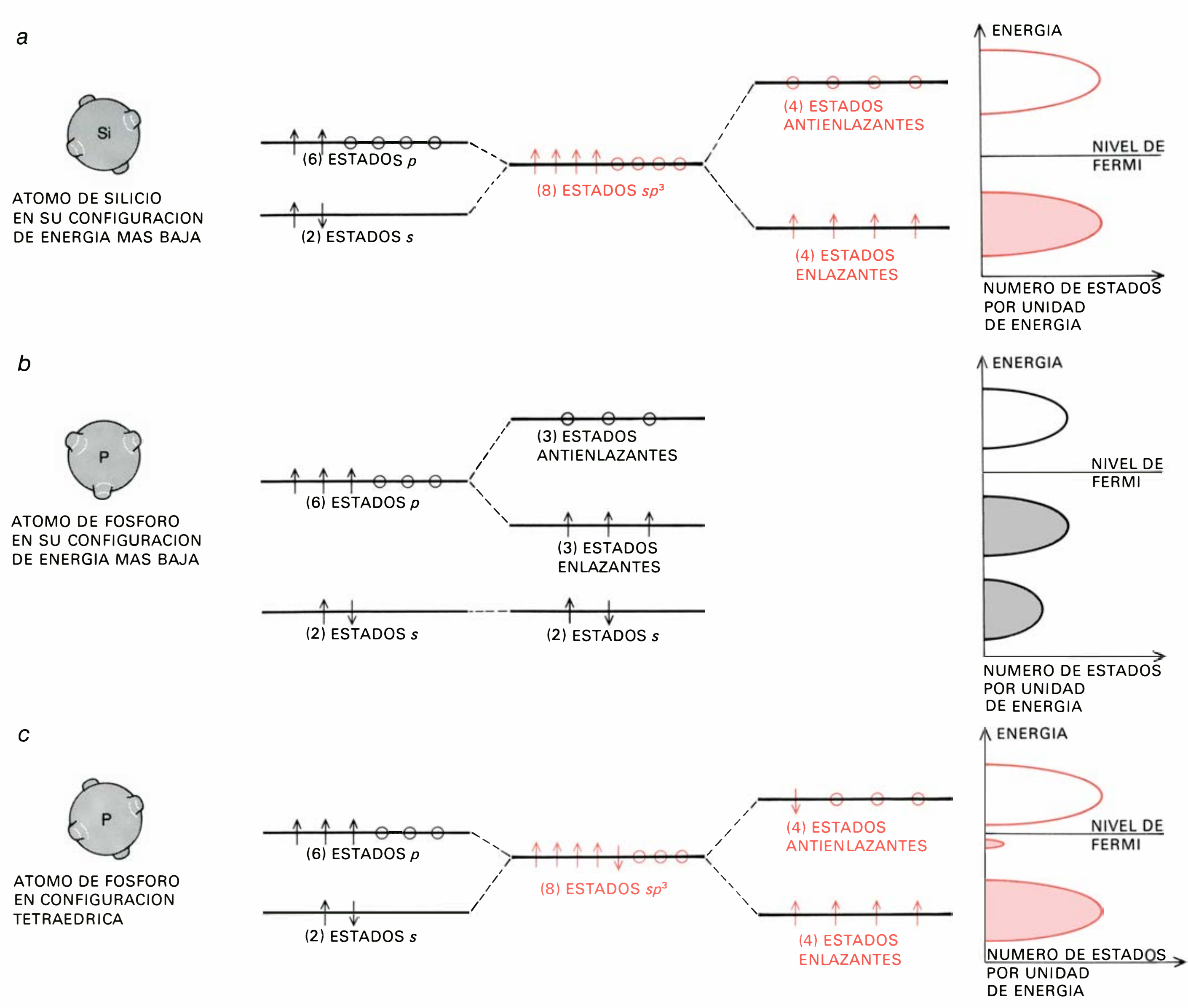
El primer paso importante en la obtención de una explicación del distinto comportamiento entre todos estos tipos de semiconductores amorfos fue dado por Marc A. Kastner, uno de mis colegas en el Instituto de Tecnología de Massachusetts. Este autor señaló que la estructura electrónica de los átomos calcógenos es fundamentalmente diferente de la estructura de todos los otros átomos que suelen estar presentes en los semiconductores amorfos, ya que poseen dos electrones externos que no participan en el enlace químico. Ovshinsky demostró cómo estos electrones podrían explicar las propiedades de conmutación reversibles de los semiconductores amorfos: los electrones en cuestión constituyen portadores libres sin exigir que se rompan los enlaces químicos responsables de la estabilidad del material. Sin embargo, otras propiedades peculiares de los vidrios calcógenos están todavía pendientes de elucidación.

Quizá la característica más insólita de los vidrios calcógenos sea su falta de una densidad medible de espines desapareados. Aun en el caso de que todos los átomos del sólido satisfagan localmente sus requisitos de enlace, cabría esperar que, a temperatura ambiente, la energía térmica disponible fuera suficiente para romper algunos de los pares de energía más alta, tal como sucede en la mayoría de las sustancias. En 1975, Philip W. Anderson, de los Laboratorios Bell, advirtió que si dos electrones con espines opuestos, próximos a un mismo átomo de un vidrio calcógeno, se atraen realmente entre sí, el estado de energía más baja no tendría densidad alguna de espines desapareados. Puesto que todos los electrones están cargados negativamente, deberían repelerse mútuamente, pero, bajo ciertas condiciones, las interacciones entre los electrones de un metal y las vibraciones de los átomos del mismo pueden conducir a una atracción entre dichos electrones y dar lugar al

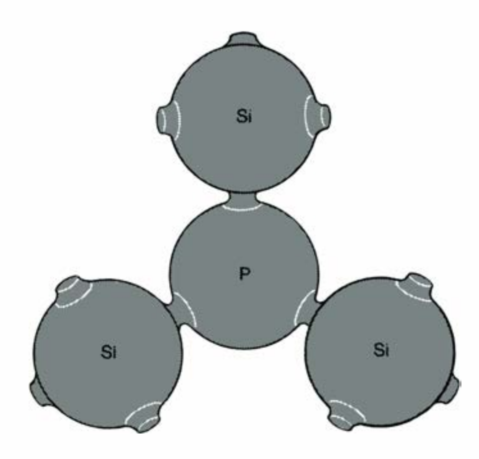
fenómeno de la superconductividad. Sin embargo, nunca se había sugerido antes la posibilidad de un efecto análogo a éste en los no metales.

Mott y sus colaboradores lanzaron la idea de que los importantes estados localizados de los vidrios calcógenos no surgen debido al desorden del material, tal como Anderson había postulado, sino que son una consecuencia de la existencia de defectos bien definidos, análogos a los que se presentan en los semiconductores cristalinos. Dichos autores sugirieron que el defecto en cuestión consistiría en un enlace móvil sobre un átomo calcógeno y demostraron además que la hipótesis de una atracción efectiva entre electrones localizados podría explicar también el peculiar comportamiento que los vidrios calcógenos muestran bajo la influencia de la luz. Ellen Yoffa y yo pudimos llegar a demostrar por qué dicho modelo da cuenta además de la relativa insensibilidad que presenta en tales vidrios la conducción respecto a la presencia de impurezas químicas. Sin embargo, subsisten dos importantes problemas pendientes de solución: ¿cómo es posible que electrones localizados de un aislador muestren una atracción efectiva? ¿Y por qué se produce este hecho en los vidrios calcógenos y no en los semiconductores amorfos tetraédricos?

Hace algunos años, Ovshinsky sugirió que las interacciones entre los pares de electrones aislados de los diferentes átomos y las interacciones con su entorno local pueden crear estados localizados en la banda de energías prohibidas o hueco energético de los vidrios calcógenos. A partir de esta hipótesis, Kastner, Fritzsche y yo pudimos identificar recientemente la naturaleza de los centros activos adecuados en los vidrios calcógenos, así como también de demos-



LA ESTRUCTURA ELECTRONICA de los semiconductores está influenciada por sus esquemas enlazantes. En el caso del silicio *(a)*, la configuración del estado de energía más baja se alcanza cuando cada átomo está rodeado por un tetraedro regular de otros átomos. Los estados *s* y *p* se hibridizan y después se desdoblan en estados enlazantes y antienlazantes. En el fósforo *(b)*, la configuración de energía más baja exige que cada átomo está enlazado con tres de sus vecinos más próximos. Puesto que los estados *s* y *p* no se hibridizan, cada uno de ellos forma una banda de energía llena en el sólido. Para que el fósforo se enlace tetraédricamente *(c)*, los estados *s* y *p* deben hibridizarse con el fin de dar lugar a cuatro estados enlazantes. Dado que se fuerza a que un electrón se sitúe en un estado antienlazante, la configuración general es más energética de lo que lo es para un enlace trigonal. De aquí que el enlace trigonal sea la configuración preferida. Sin embargo, cuando se difunden pequeñas cantidades de fósforo en silicio cristalino, las limitaciones geométricas impuestas por el cristal fuerzan a que los átomos de fósforo ocupen configuraciones tetraédricas. La energía del electrón situado en el estado antienlazante puede disminuir ligeramente, moviéndose en una órbita alrededor de los átomos de silicio vecinos, de forma tal que su energía se sitúe justo por debajo del límite inferior de la banda de conducción. Los electrones antienlazantes son excitados fácilmente a la banda de conducción por energías térmicas y, en consecuencia, la presencia de pequeñas concentraciones de fósforo aumenta fuertemente la conductividad eléctrica del silicio, elemento de la columna IV.



LA MODIFICACION QUIMICA o contaminación, que es un medio de modular la conductividad, es más difícil en un sólido amorfo que en un sólido cristalino, debido a que es posible que todos los requisitos de enlace covalente de cada átomo se satisfagan en ausencia de las limitaciones geométricas que impone un cristal periódico. Aquí, un átomo de fósforo en su estado trigonal normal, o de más baja energía, se incorpora a una matriz de silicio. Tal como puede verse en la figura de la página anterior, el hueco de energía en el fósforo ligado trigonalmente es grande y el nivel de Fermi se sitúa cerca del centro de dicho hueco. Por tanto, la adición de fósforo al silicio amorfo no altera la conductividad. Si se pretende una contaminación efectiva del silicio amorfo, debe forzarse al fósforo a enlazarse tetraédricamente, a pesar del consiguiente incremento en la energía. Esta situación se ha logrado recientemente con la ayuda de una gran concentración de hidrógeno.

trar que la presencia de átomos enlazantes con menos de cuatro vecinos próximos puede conducir automáticamente a una atracción efectiva entre los electrones localizados en dichos centros. El primer paso hacia la comprensión de las propiedades de los vidrios calcógenos consiste en señalar que el defecto eléctricamente neutro de más baja energía para un átomo calcógeno en un sólido covalente no es un enlace móvil sino un enlace adicional. Pero la presencia de cualquiera de estos dos defectos daría lugar a espines desapareados en el vidrio. Si se extrajera un electrón de un defecto y se apareara con un electrón de otro, la repulsión electrostática resultante produciría en general un aumento considerable de energía. La extracción de un electrón de un átomo calcógeno enlazado trigonalmente conduce a un estado de baja energía, debido a que el átomo resultante es similar a un átomo pnictido en su propia configuración. Por otra parte, si el electrón extraído se coloca sobre otro átomo calcógeno enlazado trigonalmente, resulta un estado de alta energía, dado que entonces existen dos electrones antienlazantes en la misma vecindad.

Sin embargo, existe otra posibilidad. El electrón adicional puede desplazarse hasta alcanzar a cualquiera de los tres átomos vecinos más próximos y el enlace entre ese átomo y el vecino ligado trigonalmente puede sencillamente romperse. Así, resulta un átomo calcógeno con un electrón extra que enlaza sólo con el vecino más próximo. Ahora bien, dicho átomo es electrónicamente similar a un halógeno (columna VII) en su configuración enlazante apropiada, y, por tanto, representa un estado de energía relativamente baja. En este caso, la transferencia de un electrón desde un centro enlazado trigonalmente a otro, con lo cual se produce la ruptura de un enlace del segundo centro, disminuye la energía del material.

Este es el origen de la atracción efectiva entre electrones localizados, sugerida por Anderson. Por otra parte, puede verse fácilmente que cada uno de los centros resultantes, alternativamente cargados positiva y negativamente, sólo tienen espines apareados, con lo cual se explica la imposibilidad de observar cualquier densidad de espines desapareados en estos vidrios. Finalmente, el modelo aclara por qué los materiales tetraédricos amorfos, tales como el silicio y el germanio amorfos, no muestran el mismo comportamiento, puesto que es imposible dar lugar a cinco enlaces únicamente con electrones *s* y *p*.

Casi al mismo tiempo en que aumentaba nuestro conocimiento de los vidrios calcógenos, se producía un considerable avance experimental en el campo de los semiconductores tetraédricos amorfos. Walter E. Spear y Peter Le Comber, de la Universidad de Dundee, fueron capaces de lograr la adulteración selectiva de una forma particular de silicio amorfo: la forma resultante de la descomposición del gas silano (SiH_4). A continuación, fabricaron diodos rectificadores con este material; este hecho tenía una gran importancia, puesto que establecía la posibilidad de desarrollar dispositivos electrónicos ordinarios basados por completo en materiales amorfos. Esta posibilidad se convirtió en realidad poco tiempo después, cuando David E. Carlson y Christopher R. Wronski, de los Laboratorios RCA, fabricaron células solares con una eficiencia de hasta un 6 por ciento partiendo de dicho material. En la actualidad, parece ser que la ruptura de la barrera económica, necesaria para alcanzar la generación en gran escala de energía eléctrica con células solares, podría muy bien alcanzarse gracias a la tecnología de los materiales amorfos, cuyo precio es intrínsecamente bajo, en lugar de llegar a través de una reducción en los costes de producción de dispositivos cristalinos uniformes.

Todos los resultados experimentales sobre los semiconductores amorfos tetraédricos pueden interpretarse en términos del modelo de centros activos descrito anteriormente, excepto en lo referente al hecho de que en estos materiales la interacción de los electrones localizados debe ser repulsiva en lugar de atractiva. Tal como se preparan las películas de semiconductores amorfos, éstos poseen densidades relativamente altas de enlaces móviles, que constituyen el defecto de energía más baja posible. Estos enlaces proporcionan los centros activos, que no sólo controlan las oportunas propiedades electrónicas, sino que también son los responsables de las elevadas densidades de espines desapareados y hacen difícil la adulteración selectiva de estos materiales. Aunque la mayoría de los enlaces móviles pueden ser eliminados mediante un proceso de recocido de las películas por debajo de su temperatura de cristalización, todavía no ha sido posible obtener mediante tratamientos térmicos una densidad lo suficientemente baja como para permitir que la conductividad eléctrica resulte fuertemente modificada por adulteración selectiva con impurezas. No obstante, el hecho de que se pueda producir tal adulteración en películas preparadas a partir de silano sugiere que el hidrógeno residual fija suficientemente dichos enlaces y, en consecuencia, permite lograr el control de la correspondiente conducción eléctrica mediante la adición de impurezas escogidas adecuadamente. Esta posibilidad ha sido confirmada en el trabajo de William Paul y sus asociados en la Universidad de Harvard, al conseguir la adulteración selectiva de películas amorfas tetraédricas exponiéndolas directamente a gas hidrógeno durante el proceso de deposición.

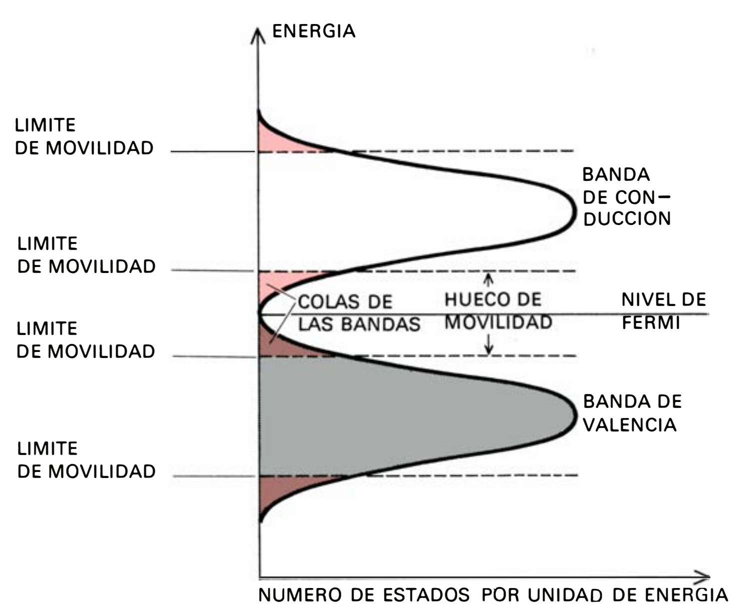
Como ya he mencionado, la conductividad de los vidrios calcógenos resulta ser bastante insensible a las impurezas. Sin embargo, el trabajo llevado a cabo en este sentido por Ovshinsky y sus colaboradores se ha visto coronado por el éxito, ya que recientemente han conseguido aumentar la conductividad de tales materiales en varios órdenes de magnitud mediante un proceso de alteración química. Estas investigaciones demuestran que es posible saturar los lugares cargados tanto positiva como negativamente en los vidrios calcógenos. Los buenos resultados logrados en la modulación de la conductividad de estos vidrios estables permiten abrir un nuevo campo de aplicaciones industriales, que van desde las

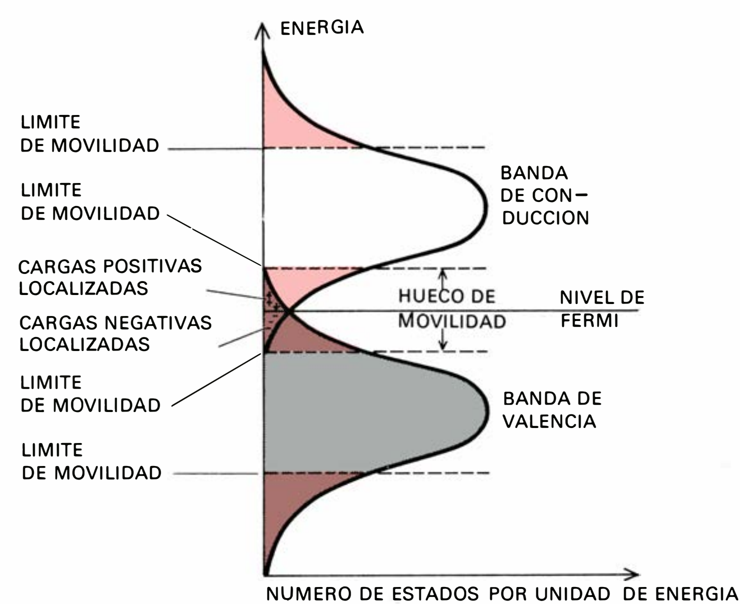
células solares de bajo coste a la producción de diferentes dispositivos de estado sólido totalmente amorfos.

El descubrimiento tardío de que las propiedades electrónicas de los semiconductores amorfos están controladas por centros activos perfectamente definidos, al igual que sucede en los semiconductores cristalinos, debería llegar a conducir a la fabricación de los análogos amorfos de muchos de los dispositivos electrónicos de los que ahora disponemos únicamente con tecnología cristalina. Por otra parte, en el caso de los vidrios calcógenos, la ausencia de una estructura rígida introduce una flexibilidad adicional que amplía enormemente sus posibilidades. Ovshinsky señaló, hace ya muchos años, que la composición continuamente variable de las aleaciones calcógenas hace posible su elaboración adaptándolas al máximo a las exigencias de una determinada aplicación. Esto ya se ha hecho, por ejemplo, modificando la banda de energías vacía y la constante dieléctrica de los vidrios calcógenos en la forma deseada. También se puede variar la densidad y naturaleza de los centros cargados positiva y negativamente de modo alternativo, modificando la composición de los vidrios mediante la utilización de diferentes procedimientos de preparación o por subsiguientes tratamientos térmicos. Todas estas perspectivas permiten esperar una gran expansión en el número de dispositivos que puedan ser fabricados a partir de estos materiales de fácil preparación.

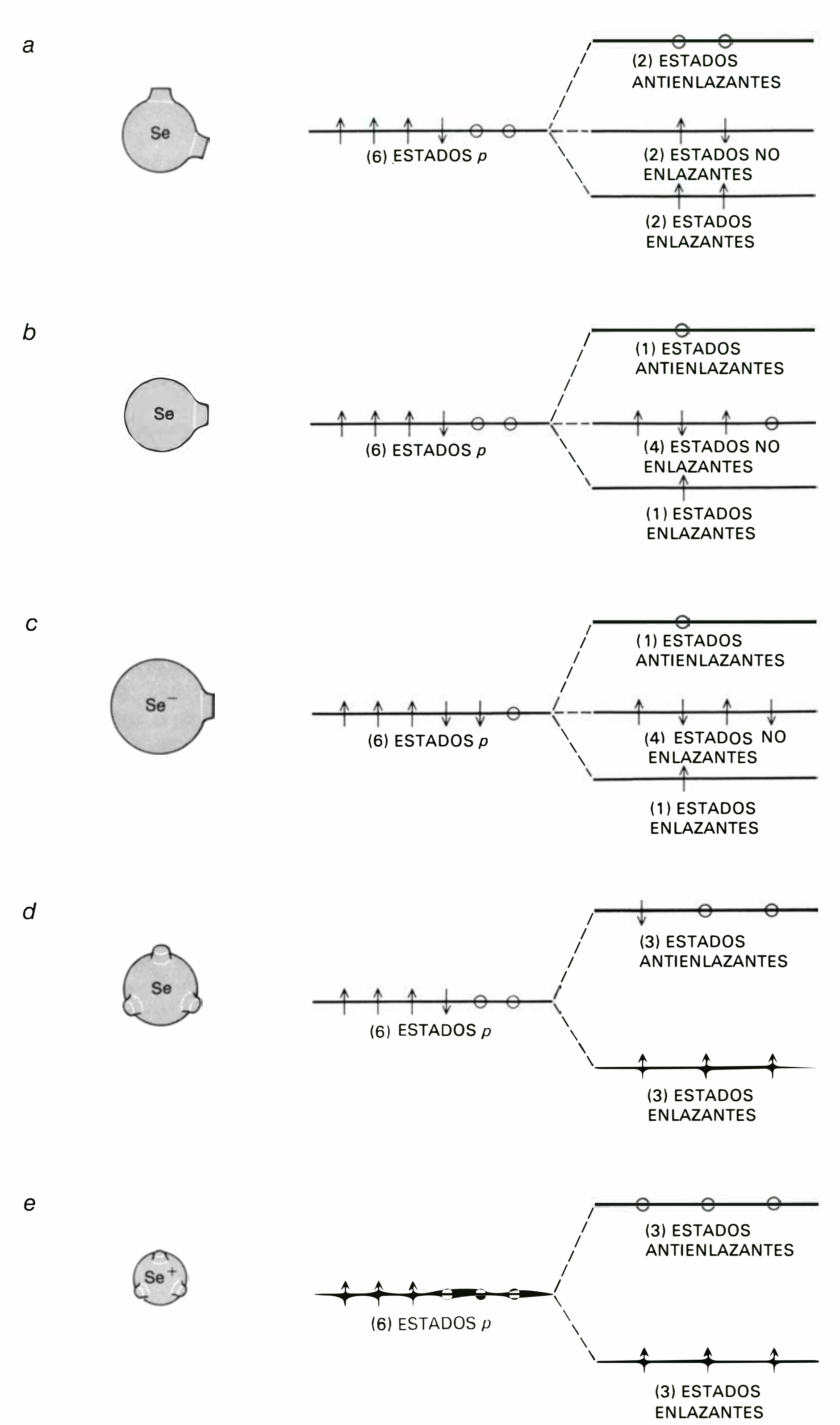
Hasta ahora, la aplicación comercial más importante de los vidrios calcógenos ha sido, sin duda, la electrofotografía (conocida también como xerografía), proceso en el cual se hace uso de las propiedades fotoconductoras de un vidrio de selenio. La superficie superior de una película de este material se suele cargar positivamente a unos 1000 voltios mediante un filamento que pulveriza iones positivos sobre ella. Como consecuencia, aparece una carga negativa "imagen" en el soporte metálico sobre el que el vidrio está depositado. La luz reflejada por el documento a copiar se hace llegar entonces hasta dicha película. En los lugares donde existe una impresión sobre el original, la luz se absorbe en mayor proporción que se refleja; la energía luminosa reflejada donde no hay nada impreso da lugar a la formación de pares electrón-agujero cerca de la superficie superior del vidrio. El campo eléctrico intenso que existe a través del vidrio separa dichos pares. Los electrones se mueven hacia arriba y neutralizan los iones positivos existentes en dicha superficie superior; los agujeros se propagan a través del vidrio y neutralizan finalmente la carga negativa existente en el soporte metálico. De este modo, la superficie superior del vidrio será eléctricamente neutra allí donde no había nada impreso en el original, pero quedará todavía cargada positivamente donde sí lo había. Acto seguido, ciertas pequeñas partículas de un colorante negro *(toner)*, cargadas negativamente, son atraídas hacia las zonas del vidrio cargadas positivamente. A continuación, el colorante se transfiere a un papel cargado positivamente. En la última fase del proceso de copia, el colorante se fija permanentemente sobre dicho papel mediante calor.

Durante el pasado año se han producido avances significativos en nuestra comprensión de los procesos electrofotográficos. Un problema persistente ha sido la inexistencia de un tiempo bien definido para que los agujeros excitados por la luz alcanzasen el soporte metálico. El reciente trabajo de Harvey Scher y sus colaboradores, en el Centro de Tecnología Joseph C. Wilson, ha demostrado que este hecho puede ser una consecuencia de cualquier proceso en el que exista una distribución de los tiempos característicos invertidos por los agujeros en alcanzar dicho soporte. Marvin Silver, de la Universidad de Carolina del Norte, ha sugerido que el proceso podría consistir en la captura continua de agujeros por los centros cargados negativamente que existen en el cristal. Más recientemente, el modelo de valencia alternada para los





LOS SEMICONDUCTORES AMORFOS que no estén fuertemente desordenados *(izquierda)* poseen bandas de valencia y de conducción similares a las de los correspondientes semiconductores cristalinos. La característica distintiva de las bandas de los sólidos amorfos es el reemplazo de los bien delimitados bordes de las bandas, típicos de los cristales, por lo que se denominan "colas de las bandas", estados localizados que se extienden a lo largo del hueco de energía. Los estados localizados se separan del resto de los estados situados en la parte principal de las bandas por "límites de movilidad". La región situada entre los límites de movilidad de las bandas de valencia y de conducción es el "hueco de movilidad". Este hueco de movilidad juega en los semiconductores amorfos el mismo papel desempeñado por el hueco de energía en los semiconductores cristalinos. La existencia de impurezas químicas o defectos en la configuración de enlaces locales puede conducir a importantes cambios estructurales en el hueco de movilidad *(no representados en la figura)*. Si el desorden es grande, tal como es propio de los vidrios multicomponentes *(derecha)*, las colas de las bandas de conducción y de valencia pueden solaparse en el hueco de movilidad. Este hecho conduce a una redistribución de la carga eléctrica a medida que los electrones se mueven desde un estado localizado a otro con el fin de disminuir su energía. El resultado es una alta densidad de colectores cargados positiva y negativamente, que disminuyen la movilidad de los portadores y hacen el material menos sensible a los esfuerzos para modular su conductividad por medios químicos, es decir, por contaminación *(doping)*.

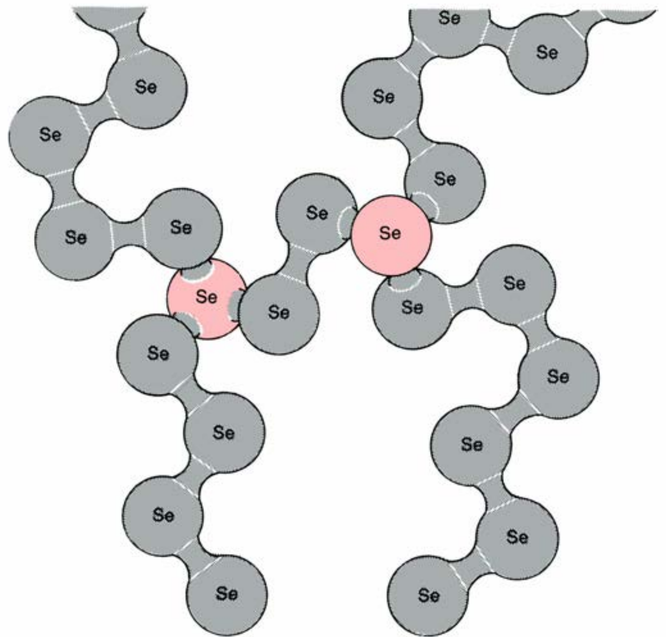


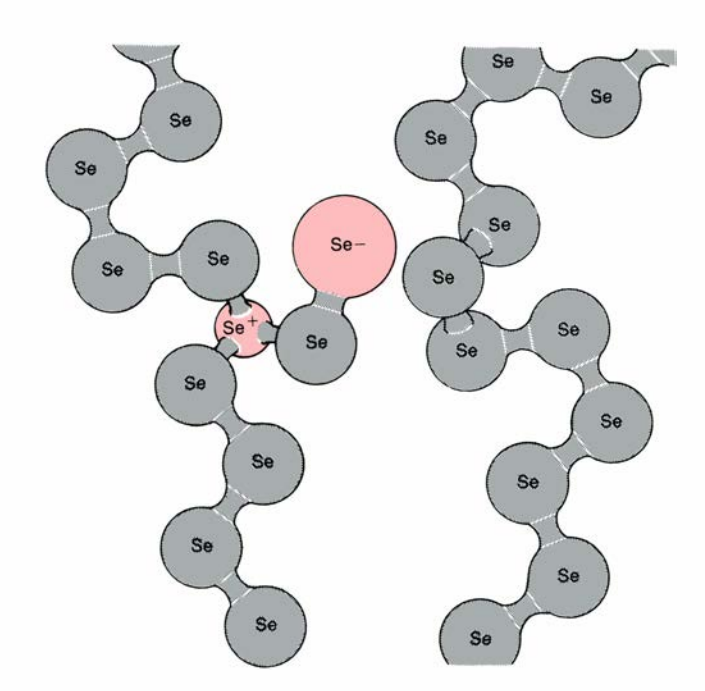
EL ATOMO DE SELENIO, uno de los elementos calcógenos, puede adoptar una amplia variedad de configuraciones enlazantes covalentemente en los vidrios calcógenos. Cuando el selenio se enlaza con dos de sus vecinos más próximos, el átomo está en la configuración de más baja energía *(a)*. El ángulo de enlace óptimo es entre 100 y 105 grados. Cuando el selenio forma sólo un enlace con un vecino *(b)*, el otro enlace queda "columpiándose" en un estado desapareado. (Los electrones enlazantes siempre aparean sus espines con el espín del vecino que comparte el enlace.) Si el átomo de selenio captura un electrón extra *(c)*, se parece a un átomo de halógeno (columna VII) en su configuración de energía más baja. En este estado de carga, sin electrones desapareados, el selenio actúa como final de una cadena. Cuando el selenio neutro está coordinado trigonalmente *(d)*, se fuerza al electrón extra a entrar en un estado antienlazante. Por tanto, la energía de esta configuración es mayor que la de un átomo de selenio con un enlace menos, pero menor que la del enlace móvil ilustrada en *b*. Si se extrae un electrón de un átomo de selenio coordinado trigonalmente *(e)*, se obtiene un estado de energía muy baja. El selenio así coordinado forma enlaces cruzados entre las cadenas. Una combinación de átomos de selenio en las configuraciones *c* y *e* posee una energía casi tan baja como la de dos átomos en la configuración *a*. En consecuencia, la combinación *c-e*, conocida como par de valencia alterna, se produce frecuentemente en los vidrios.

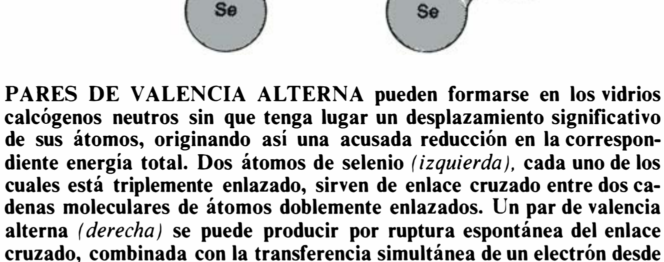
vidrios calcógenos ha clarificado diversos aspectos de la electrofotografía. En este modelo, la presencia de altas densidades de colectores cargados positiva y negativamente facilita la comprensión de cómo los agujeros pueden ser capturados continuamente. Por otra parte, la existencia de centros cargados proporciona un mecanismo sencillo para conseguir la fijación sobre la superficie del vidrio tanto de los iones positivos como de las partículas del colorante.

Las aplicaciones de las propiedades fotoconductoras de los vidrios calcógenos aumentan de continuo. La Xerox Corporation ha desarrollado una copiadora de color basada en una estructura de tres cepas de diferentes vidrios calcógenos, que responde a la luz a lo largo de todo el espectro visible. Este dispositivo se utiliza con combinaciones de filtros coloreados y partículas de diversos colorantes para hacer copias rápidas de originales en color. Muy recientemente, Hitachi y la Japan Broadcasting Corporation han desarrollado un tubo en miniatura de toma de imágenes de alta sensibilidad (un dispositivo capaz de convertir una imagen en señales eléctricas), cuyo elemento activo es una aleación de selenio, arsénico y teluro amorfos. Este tubo, conocido como "Saticon", tiene menos de dos centímetros y medio de longitud y ya se utiliza en pequeñas cámaras de televisión en color.

Todavía no he analizado las singulares propiedades de conmutación de los vidrios calcógenos, que fueron descubiertas y desarrolladas por Ovshinsky en el período comprendido entre 1958 y 1968. La publicación de los oportunos resultados despertó un enorme interés hacia todo el campo de los semiconductores amorfos, pero también dio lugar a una gran controversia. Hubo afirmaciones inmediatas en el sentido de que los dispositivos en cuestión eran irreproducibles, irreversibles, inseguros, de vida corta e inestables. En numerosos casos, la "prueba" de tales inconvenientes fue presentada en forma de datos obtenidos en experiencias que no tenían ninguna relación con el fenómeno de la conmutación. Siempre es posible "conmutar" cualquier no metal al menos una vez, en el sentido de que la aplicación de un campo eléctrico suficientemente intenso llegará a romper el carácter aislante del material, produciendo un aumento agudo de la conductancia. Este tipo de ruptura es generalmente irreversible; la conmutación propiamente dicha debe suprimir tales efectos en lugar de depender de ellos. En la actualidad, resulta eviden-







PARES DE VALENCIA ALTERNA pueden formarse en los vidrios calcógenos neutros sin que tenga lugar un desplazamiento significativo de sus átomos, originando así una acusada reducción en la correspondiente energía total. Dos átomos de selenio *(izquierda)*, cada uno de los cuales está triplemente enlazado, sirven de enlace cruzado entre dos cadenas moleculares de átomos doblemente enlazados. Un par de valencia alterna *(derecha)* se puede producir por ruptura espontánea del enlace cruzado, combinada con la transferencia simultánea de un electrón desde uno de los átomos de selenio triplemente enlazados hasta un átomo próximo a aquél por donde se haya roto el enlace cruzado. Puesto que tal transferencia electrónica disminuye la energía total del sólido, casi todos los átomos de selenio enlazados trigonalmente se convierten en miembros de un par de valencia alterna. Este hecho da lugar a consecuencias físicas muy importantes: la desaparición virtual de electrones con espines desapareados y la aparición de concentraciones grandes, e iguales entre sí, de colectores cargados positiva y negativamente.

te que la estructura electrónica de los vidrios calcógenos tiende a suprimir el efecto de ruptura al mismo tiempo que favorece la conmutación reversible real, y ello constituye precisamente la base de su superioridad cuando se aplican a los dispositivos electrónicos. Es más, estos conmutadores poseen una relación de conductancia extremadamente alta (aproximadamente de un millón) entre los estados de conducción y no conducción, un tiempo de conmutación muy pequeño (inferior a una mil millonésima de segundo) y una vida larga (algunos dispositivos han sido conmutados más de 10^{14} veces sin averías).

Resulta un tanto difícil comprender la ola de irracionalidad que siguió a la publicación del trabajo original de Ovshinsky, sobre todo en lo que se refiere a la condena de la tecnología correspondiente por parte de muchos científicos e ingenieros, que limitó seriamente el campo de los semiconductores amorfos y su utilización potencial en la producción de dispositivos electrónicos comerciales. Este hecho resulta especialmente destacable a la vista del éxito de la xerografía, un proceso en el que un semiconductor amorfo se somete a enormes variaciones del campo eléctrico, y de luz, temperatura y humedad, exigiéndose al mismo tiempo que el material cubra una gran superficie e incluso que sea capaz de resistir el agrietamiento, la descomposición y la cristalización.

Poco después de la publicación del trabajo de Ovshinsky surgieron dos auténticas controversias científicas. La más importante se refería a la naturaleza del proceso de conmutación y, en particular, acerca de si dicho proceso era fundamentalmente un fenómeno electrónico o térmico. Si fuera electrónico, el estado conductor representaría un aumento de la concentración de portadores libres (y quizá también de su movilidad), en una situación de no equilibrio. Aunque los estados de no equilibrio electrónicos ya eran bien conocidos en los sólidos cristalinos, nunca se había conseguido el mantenimiento de tales estados en los materiales homogéneos, con la caída de sólo una pequeña fracción del voltaje conmutado a través del correspondiente dispositivo. Por otra parte, otra explicación posible del estado conductor consistía en afirmar que se debía simplemente a una especie de filamento central caliente, siendo el origen de la disminución de la resistencia el acusado aumento de la conductividad que caracteriza a todos los semiconductores cuando crece la temperatura. El aumento de temperatura en este mecanismo sería simplemente una consecuencia del calentamiento de la resistencia que tiene lugar siempre que una corriente pasa a través de un medio cualquiera. El fenómeno de ruptura que se produce a continuación, denominado "fuga o disparo térmicos", es bien conocido en muchos semiconductores cristalinos y puede inducirse en otros muchos materiales si la muestra es lo suficientemente gruesa y conduce mal el calor. Aun cuando no existe una razón absoluta por la cual tal mecanismo térmico haya de originar dispositivos de menor calidad, sí es cierto que la migración de los átomos se produce más de prisa a temperaturas altas que a las bajas; de aquí que los dispositivos que operan a temperaturas elevadas tengan en general vidas operativas más cortas que los que trabajan a temperatura ambiente.

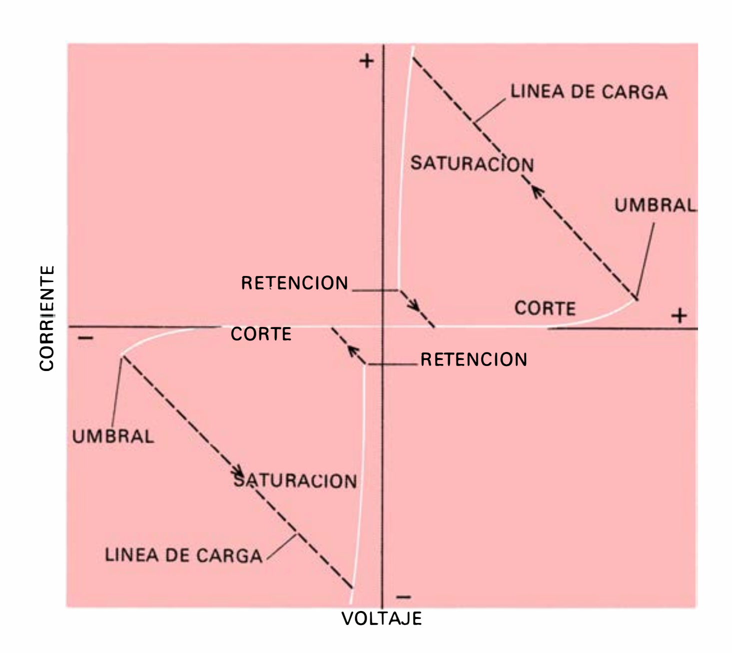
Aun cuando ya en el año 1971 H. K. Henisch y sus colaboradores de la Universidad del Estado de Pennsylvania poseían pruebas experimentales a favor de que es un mecanismo electrónico el causante del fenómeno de la conmutación en los semiconductores amorfos, esta controversia sólo ha sido resuelta definitivamente en fecha mucho más reciente. En primer lugar, Melvin P. Shaw, de la Universidad del Estado de Wayne,

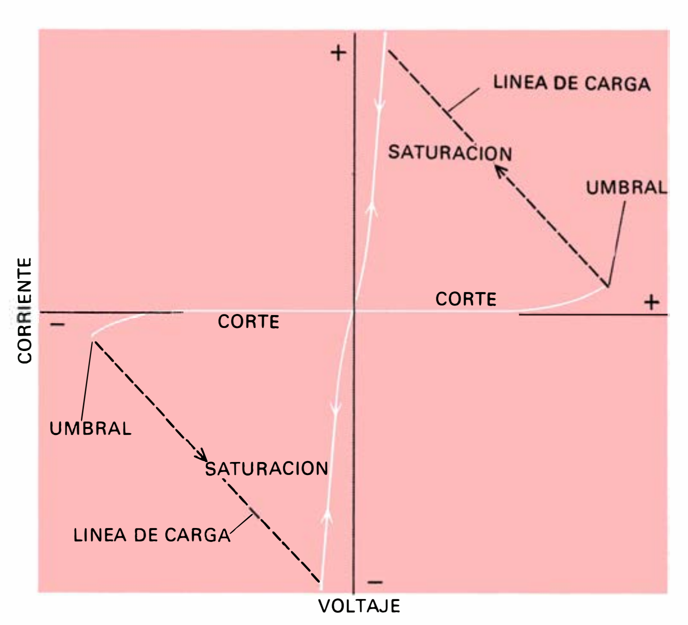
trabajando conjuntamente con Scott Holmberg y W. Derek Buckley, de la firma Energy Conversion Devices, demostraron que el estado conductor de un conmutador semiconductor amorfo podía inducirse en menos de una mil millonésima de segundo, y este tiempo es demasiado corto como para que en él pueda producirse cualquier calentamiento significativo. Este resultado indica que existe un mecanismo de conmutación electrónica, pero no rechaza la posibilidad de que en el proceso de conmutación intervenga también el ya citado filamento central caliente. Más tarde, Gary C. Vezzoli y sus colaboradores, en el Picatinny Arsenal, observaron la emisión de luminiscencia por parte del filamento conductor en el material de conmutación y encontraron que tenía el mismo carácter que la luminiscencia de un material frío excitado electrónicamente. Finalmente, Kurt E. Petersen y yo mismo procedimos a medir el tamaño de dicho filamento conductor por cuatro métodos diferentes y encontramos que es tal que la temperatura del filamento no puede aumentar por encima de 60 grados Celsius cuando el conmutador está operando bajo condiciones normales. En conclusión, parece ser que el fenómeno de la conmutación en los vidrios calcógenos responde a un mecanismo fundamentalmente electrónico, tal como Oushinsky mantuvo originalmente.

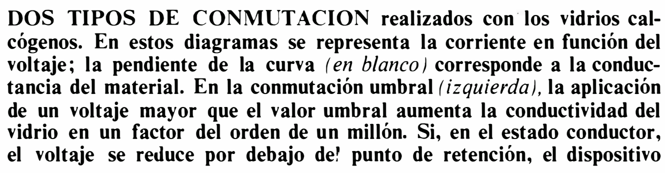
La segunda controversia se refería al fenómeno de conmutación como efecto de formación o "primer disparo". En algunos dispositivos, el primer proceso de conmutación exige un voltaje significativamente mayor que todos los procesos subsiguientes. A pesar de que ya se sabía que no todos los dispositivos muestran efectos de formación e incluso que los mejores conmutadores poseen un voltaje umbral totalmente estable, diversos grupos de trabajo mantuvieron que los efectos de formación constituyen una parte fundamental del proceso de conmutación de los semiconductores amorfos. Sin embargo, Petersen y yo hemos demostrado recientemente que los efectos de formación se presentan en los mejores conmutadores tan sólo cuando la sección transversal del dispositivo es demasiado pequeña o la corriente aplicada demasiado grande.

Las recientes experiencias de conmutación, consideradas conjuntamente con los nuevos avances en nuestra comprensión de la estructura electrónica de los semiconductores amorfos, han clasificado la singular combinación de propiedades de los vidrios calcógenos que permiten la producción de ciertos dispositivos particularmente importantes. En primer lugar, los vidrios deben contener porcentajes relativamente altos (del 30 al 70 por ciento) de átomos que se enlacen tetraédrica o trigonalmente, con lo cual se entrecruzan las cadenas de los átomos calcógenos. Este hecho da lugar a una estabilidad estructural que retarda la cristalización del vidrio. En segundo lugar, las composiciones particulares deben ser tales que la conductividad eléctrica de los dispositivos a temperaturas ambientales sea extremadamente baja. Si la conductividad es baja, el calentamiento de la resistencia es mínimo, y este hecho puede originar un efecto de fuga térmica. Finalmente, es importante tener una concentración bastante alta de átomos calcógenos, con el fin de proporcionar un número suficientemente grande de colectores cargados positiva y negativamente, necesarios para retardar una ruptura eléctrica irreversible del material.

Una vez que los resultados experimentales quedaron bien establecidos, se empezó a comprender el mecanismo detallado de la conmutación en los semiconductores amorfos. Shaw y sus colaboradores desarrollaron un modelo de circuito que explica satisfactoriamente el comportamiento transitorio de la corriente con relación al voltaje. Petersen y yo propusimos un modelo electrónico cuantitativo para la recuperación del






DOS TIPOS DE CONMUTACION realizados con los vidrios calcógenos. En estos diagramas se representa la corriente en función del voltaje; la pendiente de la curva *(en blanco)* corresponde a la conductancia del material. En la conmutación umbral *(izquierda)*, la aplicación de un voltaje mayor que el valor umbral aumenta la conductividad del vidrio en un factor del orden de un millón. Si, en el estado conductor, el voltaje se reduce por debajo de! punto de retención, el dispositivo conmuta al estado de baja conductancia o de corte. La conmutación de memoria *(derecha)* requiere un vidrio con propiedades diferentes. Si el vidrio se mantiene durante un cierto tiempo en el estado de alta conductancia, sigue siendo conductor aun después de haber eliminado el voltaje aplicado. En este estado, debido a un filamento de material cristalino, puede restablecerse el estado de corte aplicando un impulso de corriente adecuado, cuya acción disuelve el filamento y restituye el vidrio.

estado de baja conductividad una vez que se ha anulado el voltaje a lo largo del dispositivo. Las predicciones de este modelo se mostraron en buen acuerdo con las medidas subsiguientes. Por otra parte, una consideración detallada de la especial naturaleza de la estructura electrónica de los vidrios calcógenos ha conducido al desarrollo de un modelo preliminar para el propio mecanismo de la conmutación. En este enfoque, el estado conductor se alcanza únicamente después de que los colectores positiva y negativamente cargados que existen generalmente en el vidrio se llenan con los portadores excitados por el campo eléctrico aplicado exteriormente. Cuando todos los colectores están llenos, la vida de un portador inyectado aumenta bruscamente desde un valor muy inferior al tiempo que el portador necesita para atravesar todo el espesor de la lámina, hasta un valor mayor que el tiempo que dicho portador necesita para atravesar esa lámina, produciendo así una caída en el voltaje y el correspondiente aumento de la corriente que caracteriza al estado conductor. Este modelo está de acuerdo con las recientes experiencias que indican que la movilidad de los portadores en el estado altamente conductor es la misma que la que se pone de manifiesto entre los sucesos colectores en el estado de baja conductividad. Los tiempos de respuesta, tanto para el llenado como para el vaciado de los colectores, están de acuerdo con las predicciones del modelo de alteración de valencia de los vidrios calcógenos. En la actualidad, el proceso de conmutación se considera tan bien conocido como la mayoría de los efectos de no equilibrio en los semiconductores cristalinos.

En su trabajo original, Ovshinsky anunció también el descubrimiento de un segundo tipo de fenómeno de conmutación, en el cual el estado conductor subsiste incluso tras la desaparición total del voltaje aplicado. El estado no conductor puede ser restablecido tan sólo mediante la aplicación de un fuerte impulso de corriente. Este comportamiento, conocido actualmente como conmutación de memoria, es propio de los vidrios calcógenos cuya estructura está estabilizada por sólo una pequeña cantidad de enlaces cruzados, con lo cual se facilita considerablemente la oportuna cristalización. Ovshinsky sugirió que la conmutación de memoria es, por tanto, consecuencia de una transición entre el estado amorfo y el cristalino, fenómeno que en estos vidrios es intrínsecamente reversible. En realidad, los estudios estructurales detallados llevados a cabo por Arthur I. Bienenstock, de la Universidad de Stanford, Simon C. Moss y otros investigadores demostraron que en la clase de materiales utilizados generalmente en los conmutadores de memoria (basados en ciertas aleaciones de teluro-germanio), las elevadas condiciones de no equilibrio del umbral de conmutación inducen la formación de pequeñas zonas cristalinas de teluro semimetálico, que son las responsables de la memoria del estado conductor. Los estudios de resonancia magnética nuclear han demostrado claramente la reversibilidad intrínseca de la transición de memoria. De hecho, el estado conductor es similar al de un metal cuando se comporta como superconductor a bajas temperaturas. Se ha desarrollado un modelo cuantitativo del proceso de conmutación cuyo acuerdo con los datos experimentales es excelente.

Quizá como consecuencia del hecho de que el fenómeno de la conmutación de memoria fuera comprendido relativamente pronto, los dispositivos basados en él se desarrollaron muy rápidamente. A finales de 1970 fue introducida en el mercado por la empresa Energy Conversion Devices una memoria estable, electrónicamente alterable, de 256 bits (memoria "Ovonic"). Conocida también como memoria preferentemente de lectura (*read-mostly memory* [RMM]), este dispositivo se construye ahora por la Burroughs Corporation. La memoria Ovonic desempeña actualmente un papel de gran importancia en la tecnología de las memorias bipolares, llenando la enorme laguna existente entre las memorias de acceso aleatorio (*random-access memory* [RAM]), que son transitorias y alterables, y las memorias exclusivamente de lectura (*read-only memory* [ROM]), que son estables e inalterables. Por otra parte, dicha memoria es directamente compatible con la lógica utilizada corrientemente en los sistemas de memoria de los ordenadores, es resistente a las radiaciones ionizantes y a otras tensiones ambientales, presenta un tiempo de acceso de 50.000 millonésimas de segundo y tiene una vida de escritura/borrado de más de 100 millones de ciclos por bit. En los últimos cinco años, la memoria Ovonic ha sido mejorada continuamente, en particular en lo referente a su vida, corriente de restitución, densidad de empaquetamiento e intervalo operativo máximo. Las "curvas de aprendizaje" han sido comparables o incluso mejores que las obtenidas en otras áreas de la industria de los semiconductores.

Otro campo de posible aplicación de la conmutación de memoria lo tenemos en las memorias ultragrandes, incluyendo las utilizadas en el almacenamiento de archivos. Peter Klose, Julius Feinleib y Ovshinsky demostraron que los dos estados de memoria son ópticamente diferentes; cuando el material está en un estado es mucho más transparente a la luz que cuando está en el otro. Esto significa que, utilizando luz, puede conocerse en qué estado se encuentra el material. Es más, se observó que no sólo la lectura, sino también la escritura y el borrado pueden llevarse a cabo por los mismos medios ópticos. Puesto que la luz visible se puede focalizar sobre una región cuyo diámetro llega a ser del orden de la micra, existe la posibilidad de utilizar estos materiales de conmutación de memoria para fabricar memorias cuya capacidad sea de un billón de bits por metro cuadrado.

Aunque inicialmente no se comprendía la brevedad del tiempo de escritura/borrado (aproximadamente una millonésima de segundo), así como la simetría del ciclo de escritura/borrado, pronto se desarrolló un modelo capaz de explicar satisfactoriamente estos resultados. La conmutación óptica de memoria es otro ejemplo de una transición reversible amorfo-cristalino; la rapidez del tiempo de escritura tiene su origen en el reforzamiento de la cristalización por la luz, y la simetría del proceso de escritura/borrado es una consecuencia del hecho de que las zonas vítreas son más transparentes que las cristalinas. Tras el trabajo inicial, se desarrolló un proceso alternativo que no dependía de la cristalización. Este proceso proporcionó una mayor resolución, un tiempo de escritura todavía más pequeño y una gran reducción en la energía necesaria para la escritura. Este proceso alternativo ha sido explicado en términos de la formación reversible de burbujas de vapor en el vidrio por la acción de la luz.

Otra posibilidad diferente para lograr sistemas de memoria muy grandes se basa en la utilización de haces de electrones en los correspondientes procesos de escritura y lectura. En el modo de escritura, los estados amorfos y cristalinos de los vidrios de conmutación de memoria difieren profundamente en cuanto a la emisión de electrones secundarios al ser bombardeados por un haz de electrones. Gracias a las técnicas de que dispone y

va a disponer la óptica electrónica, será posible conseguir una región de impacto cuyo tamaño sea de 10 nanómetros, lo cual equivale a la centésima parte del tamaño mínimo que se puede alcanzar utilizando medios ópticos de luz visible.

El hecho de que los vidrios calcógenos puedan ser puestos en un estado conductor permanente mediante la luz sugiere otra importante aplicación en el campo de la impresión. Como saben todos los usuarios de las copiadoras electrostáticas, la obtención de varias copias exige varias aplicaciones de luz. No obstante, los vidrios de conmutación de memoria pueden proporcionar una placa permanente, a partir de la cual se pueden realizar un número ilimitado de copias electrofotográficas sin necesidad de exposiciones adicionales. La técnica parece prometedora en el campo de la impresión rápida a gran escala.

Uno de los sectores actuales de desarrollo más prometedor de los semiconductores amorfos es el de la reproducción en general. Tal como sucede con la conmutación de memoria, la base de muchos de los sistemas de reproducción la tenemos en la transición amorfo-cristalino. Dada la importancia que en el campo de la reproducción tiene la gradación de tonos, son características fundamentales el tamaño de los granos del material y la fracción del mismo que está cristalizado. También en este sentido las ideas originales se deben a Ovshinsky. Se puede producir un alto contraste o tonos continuos, según el tipo de aplicación. El hecho de que la imagen se pueda obtener directamente tras la exposición o bien quede en forma de imagen latente, que puede ser revelada posteriormente por calor o radiación, constituye un factor de flexibilidad adicional. En ningún caso es necesaria una fase de fijado, puesto que tanto el estado amorfo como el cristalizado son totalmente estables a las temperaturas ordinarias. Por lo demás, el mecanismo permite las adiciones o las sobreimpresiones sin necesidad de ningún género de borrado. Otra característica adicional consiste en que ciertos cambios en las propiedades físicas pueden detectarse y utilizarse con fines de reproducción. Por ejemplo, los dos estados en cuestión difieren en su humectabilidad: el material cristalino acepta el agua y rechaza la tinta aceitosa, en tanto que el material amorfo rechaza el agua y admite la tinta. Por este método se puede obtener una reproducción fotográfica a partir de la película expuesta.

El concepto amorfo-cristalino se ha extendido al uso de los organotelúridos como elementos activos. En este caso, la exposición a la luz conduce a la formación de un vidrio calcógeno particular, que forma la imagen latente. Cuando la película se calienta por encima de la temperatura a la cual el vidrio se funde, pequeñas regiones cristalinas de teluro precipitan fuera del vidrio y dan lugar así a la formación de la correspondiente imagen. Parece posible, pues, un concepto nuevo de fotografía instantánea.

Otros dos procesos característicos de ciertos vidrios calcógenos e inducidos por la luz son el fotoennegrecimiento y la fotocontaminación. En el fotoennegrecimiento, los cambios estructurales reversibles o irreversibles inducidos por la absorción de la luz conducen a una disminución significativa de la transparencia óptica del material. Este hecho sugiere muchas y muy importantes aplicaciones, desde la modulación de la luz al almacenamiento de información holográfica.

En el proceso de fotocontaminación, la difusión de ciertos metales en los vidrios calcógenos *(doping)* resulta intensificada por efecto de la luz. En general, los vidrios calcógenos se disuelven rápidamente en soluciones alcalinas, pero si contienen bastante plata son prácticamente insolubles. Este efecto, explorado por Akira Yoshikawa y sus colaboradores de la Nippon Telegraph and Telephone Public Corporation, puede ser utilizado muy ventajosamente en la industria de los semiconductores, donde la fotolitografía tiene una importancia vital en la fabricación de circuitos integrados.

Existen otros efectos estructurales en las películas que pueden ser utilizados para la formación de imágenes. La firma Energy Conversion Devices ha desarrollado una amplia variedad de materiales destinados a responder a la energía en diversas formas, que van desde cambios sutiles, tales como la alteración de la longitud de las cadenas de átomos calcógenos por causa de la luz, a variaciones en las propiedades de flujo. Algunos de estos efectos estructurales no se limitan a las películas amorfas, sino que pueden producirse también en películas cristalinas. Todos estos sistemas de formación de imágenes están caracterizados por una gran flexibilidad en la respuesta, un proceso de revelado en seco, alta resolución y excelente estabilidad.

Además de los dispositivos basados totalmente en sólidos amorfos, recientemente se han desarrollado estructuras electrónicas híbridas consistentes en heterouniones amorfo/cristalino. En 1974, Don Reinhard, Floyd O. Arntz y yo desarrollamos un modelo de bandas de energía que explica satisfactoriamente las características electrónicas y fotoconductoras de tales heterouniones. Más tarde, Petersen y yo encontramos que el modelo podía extenderse al caso en el que el semiconductor amorfo es conmutado al estado conductor.

Las heterouniones híbridas pueden ser extremadamente útiles. Por ejemplo, se puede conseguir muy fácilmente que actúen como rectificadores; su capacidad de rectificación subsiste aun después de que el vidrio sea conmutado. Por otra parte, puesto que los sólidos amorfos no poseen una estructura cristalina rígida, la densidad de estados energéticos a lo largo de la interfase entre los materiales de la heterounión en los dispositivos híbridos es significativamente menor que la propia de las heterouniones totalmente cristalinas. Puesto que los estados de la interfase tienden a degradar las cualidades de las heterouniones, su acusada reducción en los dispositivos de vidrios híbridos conduce a un comportamiento casi ideal. Petersen y yo encontramos también que el vidrio calcógeno en su estado conductor puede ser utilizado para inyectar electrones de alta energía directamente en la banda de conducción del semiconductor cristalino.

Este último descubrimiento sugirió pronto la posibilidad de que un vidrio calcógeno pudiera desempeñar el papel de emisor en un nuevo tipo de transistor. Recientemente, Petersen, Shaw y yo fabricamos tal dispositivo, utilizando silicio cristalino como base y colector. Cuando el emisor de este transistor está en el estado resistivo, las ganancias de corriente son del orden de 0,05. Sin embargo, cuando el vidrio emisor se conmuta mediante un pequeño impulso, la ganancia aumenta en un factor del orden de 200, alcanzando un valor del orden de 10. Dependiendo de la relación entre el voltaje del colector y el voltaje de la base, el transistor puede ser dejado en su estado de alta ganancia o devuelto al estado de baja ganancia. En el estado de baja ganancia, el dispositivo simple actúa como un amplificador umbral; en el estado de alta ganancia es un amplificador de enganche umbral (este sistema requiere habitualmente tres subunidades: un detector umbral, un circuito de enganche y un amplificador). El dispositivo en cuestión puede utilizarse también para convertir un sistema de os-

cilador a amplificador. Por otra parte, puesto que dicho dispositivo proporciona un estado adicional (la condición de baja ganancia), que es claramente diferente de las condiciones de saturación y corte características de un transistor ordinario, puede ser utilizado como base de una lógica ternaria. Tal lógica puede especificar tres condiciones: menos, cero y más. En consecuencia, este dispositivo híbrido constituye el medio más eficiente que se conoce para almacenar y transmitir información.

El trabajo sobre las heterouniones entre vidrios calcógenos del tipo umbral y los semiconductores cristalinos diferentes del silicio y del germanio está sólo en sus comienzos, pero no hay duda alguna acerca de la importancia de estos vidrios como eficientes inyectores de electrones. Ninguno de los materiales cristalinos con que se cuenta hoy en día puede realizar esta función.

Los dispositivos de semiconductores amorfos se encuentran en la actualidad en una fase crítica de desarrollo. A lo largo del año pasado el progreso ha sido extraordinario. Parece ser que estamos empezando a comprender las razones básicas que explican las singulares propiedades de los vidrios calcógenos; al mismo tiempo, hemos sido capaces de modular las propiedades electrónicas tanto de los vidrios calcógenos como de los semiconductores tetraédricos amorfos y podemos explicar cuantitativamente tanto la naturaleza del estado conductor como el proceso de recuperación de los conmutadores umbrales. En el pasado, el desarrollo en el conocimiento de la física inherente a un fenómeno dado ha conducido siempre a un período de rápido avance tecnológico. Un ejemplo importante lo tenemos precisamente en el campo de los transistores cristalinos. Los dispositivos originales fueron transistores de contacto puntual, cuya teoría era un misterio. Una vez desarrollada la teoría de los semiconductores, se elaboró un modelo cuantitativo para el transistor de unión, aunque todavía no se había fabricado tal dispositivo. Debido al hecho de que la física de los transistores de contacto puntual permanecía todavía oscura, en un período de dos años los transistores de unión dominaron completamente el campo. Aunque no se puede predecir el futuro con ninguna certidumbre, parece plausible la conclusión de que en el curso de los próximos años los dispositivos de semiconductores amorfos se desarrollarán muy rápidamente y serán muy ampliamente aceptados.

Ciencia y sociedad

La revolución cuántica

Hace cincuenta años, en junio o quizá julio de 1927, apareció el cuaderno III-IV del volumen 43 de la prestigiosa revista alemana de física *Zeitschrift für Physik.* En sus 150 páginas se publicaban ocho artículos, dedicados fundamentalmente a temas experimentales: espectros de absorción, fenómenos de fluorescencia, medición energética de rayos X, difusión de corpúsculos en fluidos... Por contraste, uno de esos artículos, el tercero, más que teórico era filosófico. Su título alemán, literalmente traducido, decía "Sobre el contenido intuitivo de la cinemática y mecánica teórico-cuánticas". Lo firmaba W. Heisenberg, de Copenhague. Las veintisiete páginas de ese artículo dejarían huella en el pensamiento occidental. Fundamentan la llamada interpretación de Copenhague de la mecánica cuántica. Pero, sobre todo, proporcionan un primer horizonte conceptual, que destaca el sentido de los recientes cambios de formalismo de la física atómica y los eleva a la categoría de revolución científica, la revolución cuántica de la física.

Hacía ya dos años que en volúmenes anteriores de esta misma revista Werner Heisenberg, joven profesor en el departamento de la Universidad de Gotinga dirigido por Max Born, había creado esa "cinemática y mecánica teórico-cuánticas". Había publicado en efecto un primer artículo, calificado de "místico", en el que, sin conocer la noción matemática de matriz, introducía para los observables físicos del átomo cantidades matriciales, con producto matricial. Debido a una enfermedad de Heisenberg y a la prolongada ausencia por convalecencia, son su jefe de departamento Max Born y su compañero Pascual Jordan –más cultivado éste en teoría de matrices– quienes publicaron en la misma revista otro artículo sobre "Mecánica cuántica", en que interpretaban matemáticamente el formalismo matricial de Heisenberg, y en concreto su "principio de cuantificación fuerte"

$$pq - qp = \frac{h}{2\pi i} I \qquad (1)$$

donde q y p son las matrices que corresponden respectivamente a los observables de "posición" y "momento" (la "cantidad de movimiento" newtoniana, que se define pre-relativísticamente como el producto de la masa del móvil por su velocidad). La expresión matricial (1) deja claro que el producto de las matrices p y q no es conmutativo; pq no es igual que qp, sino que su diferencia es una fracción imaginaria de la pequeñísima constante de Planck h, multiplicada por la matriz unidad, I. Una vez repuesto Heisenberg de su enfermedad, "los tres hombres", es decir, Born, Heisenberg y Jordan, publicaron en la misma revista un nuevo artículo "Mecánica cuántica II", en el que se completaba ese formalismo matricial, y se aplicaba a ejemplos más complejos. Mientras tanto, en otras publicaciones habían ido apareciendo una serie de comunicaciones de Erwin Schrödinger sobre mecánica cuántica ondulatoria y su equivalencia con la mecánica cuántica matricial de "los tres hombres", y unos trabajos de Paul A. M. Dirac sobre la llamada "teoría de transformaciones", con la que quedará clara la equivalencia matemática de ambas formulaciones cuánticas. Por otra parte, en la misma *Zeitschrift für Physik* había aparecido ulteriormente un artículo de Wolfgang Pauli que, con la nueva mecánica cuántica, resolvía elegantemente el problema de las líneas espectrales del hidrógeno, y un par de artículos de Max Born que atacaban con la nueva mecánica problemas de colisión entre partículas subatómicas.

Todos esos trabajos tenían ya más de un año, y habían implantado ese nuevo formalismo matemático que iba aplicándose a diversos campos de la física atómica. Pero a esa nueva física cuántica le faltaba la correspondiente elaboración conceptual. Esta era la preocupación de Heisenberg, que durante el curso 1926-1927 trabajaba en el Instituto de Física Teórica de Copenhague como ayudante de Niels Bohr y vivía en un estudio anexo al mismo Instituto. Bohr, famoso ya como creador del modelo cuántico de átomo (1913), tenía temperamento filosófico y era sin duda uno de los grandes pensadores de las nuevas concepciones cuánticas. Parece que el estudio en que vivía Heisenberg fue testigo de prolongadas discusiones entre los dos pensadores, en las que sin duda fueron madurando esas ideas revolucionarias. Pero tales ideas no eran compartidas en modo alguno por Schrödinger, quien estaba también aquel curso en el Instituto de Copenhague explicando su mecánica ondulatoria, y al oír de ellas "se arrepentía de haber intervenido en ese asunto". Por el contrario, Pauli protagonizó, entre bastidores, esa elaboración conceptual. Con él había mantenido Heisenberg, durante ese curso, un profundo intercambio epistolar, y de él había recibido entusiasmo y aliento para publicar sus ideas. Pero en realidad Heisenberg había redactado solo, durante unas vacaciones de esquí de Niels Bohr, el artículo que ahora aparecía en *Zeitschrift der Physik.*

En la introducción del artículo, Heisenberg retrata bien la crisis conceptual de la mecánica cuántica en 1927: "La interpretación intuitiva de la mecánica cuántica está aún llena de contradicciones internas, que se traducen en la lucha de opiniones sobre la teoría del discontinuo y del continuo, de corpúsculos y ondas. De ahí podría ya deducirse que una interpretación de la mecánica cuántica es absolutamente imposible con los conceptos cinemáticos y dinámicos ordinarios. La mecánica cuántica nació precisamente del intento de romper con los conceptos cinemáticos ordinarios y poner en su lugar relaciones entre los valores numéricos concretos que da la experiencia. Puesto que parece que esto se ha logrado, el esquema matemático de la mecánica cuántica no necesita por su parte ninguna revisión... Pero la necesidad de una revisión de los conceptos cinemáticos y mecánicos parece seguirse inmediatamente de las ecuaciones de la mecánica cuántica".

Para esta revisión de conceptos, Heisenberg tenía ante los ojos el ejemplo de otra revolución conceptual de la física, la relatividad einsteniana. Albert Einstein había introducido su relatividad especial con una crítica del concepto de simultaneidad, y con ello había ligado indisolublemente nuestros conceptos vulgarmente independientes de espacio y tiempo. En la relatividad general, el concepto de materia aparece además como curvatura geométrica del espacio-tiempo. Inspirado por este ejemplo, Heisenberg nos da su concepción epistemológica de lo que significa para él "entender intuitivamente una teoría física".

Comienza así la introducción de su artículo: "Una teoría física, a nuestro juicio, la entendemos intuitivamente cuando podemos pensar cualitativamente las consecuencias experimentales de esta teoría en todos los casos sencillos y al mismo tiempo comprobamos que la aplicación de la teoría no encierra ninguna contradicción interna. A nuestro juicio, entendemos intuitivamente por ejemplo la representación einsteniana del espacio tridimensional cerrado, por-

que las consecuencias experimentales de esta representación podemos pensarlas libres de toda contradicción. Estas consecuencias contradicen por ejemplo nuestros conceptos vulgares intuitivos de espacio y tiempo. Pero podemos convencernos de que la posibilidad de aplicar estos conceptos ordinarios de espacio y tiempo a espacios muy grandes no se sigue ni de las leyes de nuestro pensamiento ni de la experiencia".

Para realizar, dentro de ese cuadro epistemológico, la nueva revisión conceptual, Heisenberg se fija en el principio fundamental de cuantificación que hemos escrito en la ecuación matricial (1). En ella se conjuga el concepto de "posición" con los de "velocidad" y "masa", y esto le hace ya sospechar un cambio en la relación de aquellos conceptos, al tratarse de masas pequeñas y discontinuidades apreciables. Heisenberg ilustra sus ideas con dos figuras y un comentario pedagógicamente insuperables. Los hemos reproducido en su versión original. La traducción literal del comentario es la siguiente: "Imaginemos por ejemplo un punto dotado de masa que se mueve en una sola dimensión. En una teoría del continuo *(a)* se puede trazar una curva horaria *x(t)* para la trayectoria de la partícula (o con toda precisión, de su centro de gravedad), y la tangente a esta curva da la velocidad en un momento cualquiera. Por el contrario, en una teoría del discontinuo en lugar de esta curva se tiene algo así como una serie de puntos a distancias finitas *(b)*. En este caso carece evidentemente de sentido hablar de velocidad para una posición determinada, porque la velocidad sólo puede definirse mediante dos posiciones, e, inversamente, por consiguiente, a cada punto corresponden dos velocidades distintas".

Concluye así su introducción, justificando el análisis conceptual de los términos cinemáticos o dinámicos de "posición", "trayectoria", "velocidad" y "energía", análisis que realizará en la primera de las cuatro partes de su artículo. En ella adopta su postura de discreto positivismo, fundada sobre la observación física: "Si se quiere aclarar lo que se entiende por posición de un objeto, por ejemplo de un electrón..., se han de proponer ciertas experiencias mediante las cuales se pretende medir la posición del electrón; si no fuera así, ese término carecería de sentido. No faltan tales experiencias que permiten en principio determinar la posición del electrón con la precisión que se desee. Se puede, por ejemplo, iluminar el electrón y mirarlo con un microscopio".

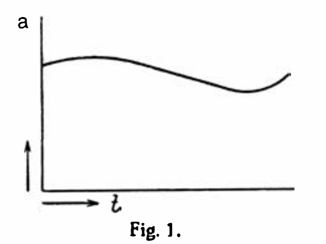


Fig. 1.

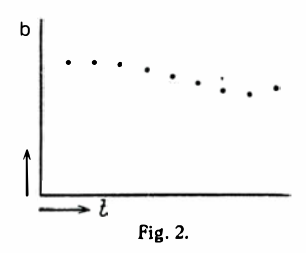


Fig. 2.

plausibel: Denkt man z. B. an die eindimensionale Bewegung eines Massenpunktes, so wird man in einer Kontinuumstheorie eine Bahnkurve $x(t)$ für die Bahn des Teilchens (genauer: dessen Schwerpunktes) zeichnen können (Fig. 1), die Tangente gibt jeweils die Geschwindigkeit. In einer Diskontinuumstheorie dagegen wird etwa an Stelle dieser Kurve eine Reihe von Punkten endlichen Abstandes treten (Fig. 2).

Según Heisenberg, en la nueva física del discontinuo (derecha), *si se conoce la posición de un móvil se desconoce su velocidad (no hay curva horaria ni tangente a ella,* izquierda).

Para realizar esa experiencia con una precisión capaz de apreciar distancias subatómicas se requiere iluminar con luz de longitud de onda extremadamente corta, es decir, con rayos gamma. La colisión de uno de esos cuantos de luz con el electrón constituye el llamado "efecto Compton", cuya cinemática había sido bien estudiada por A. H. Compton un par de años antes. En esa colisión el cuanto de luz y el electrón se intercambian energía y momento. Así, en el instante en que se determina la posición del electrón, se modifican imprevisiblemente su momento y su velocidad, y tanto más se modifican cuanto más precisa sea la determinación de la posición. Aquí es donde Heisenberg introduce su célebre relación de incertidumbre: "Sea q_1 la precisión con la que se conoce el valor de la posición q (q_1 es por ejemplo el error medio de q), es decir, en nuestro caso la longitud de onda de la luz, y sea p_1 la precisión con la que puede determinarse el valor del momento p, es decir, en nuestro caso el cambio imprevisible de p por efecto *Compton;* resulta entonces, a partir de fórmulas elementales del efecto *Compton,* que p y q están en la relación:

$$p_1\, q_1 \sim h \qquad (2)$$

Mostraremos más adelante que esta relación (2) está matemáticamente ligada de modo directo con la relación de conmutación (1)".

Admitida como esencial a la teoría cuántica esa incertidumbre relativa en las medidas de posición y momento, que queda acotada por la pequeña constante de Planck h, Heisenberg analiza el concepto de "trayectoria" como órbita bien definida del electrón en torno al núcleo atómico. Concluye que ese término "no tiene ningún sentido razonable", ya que se puede hacer una medida precisa de posición del electrón iluminándole con luz conveniente, pero en esa medida el cuanto de luz "lanza al electrón completamente fuera de su órbita". Esa órbita expresa exclusivamente una mayor probabilidad para esa medida única de la posición del electrón, probabilidad de posición que puede determinarse estadísticamente para un conjunto de átomos preparados en el mismo estado. Así quedan difuminadas en una simple nube de probabilidad las precisas órbitas cuantificadas del modelo atómico de Bohr.

Analiza después Heisenberg el concepto de "energía", cuyas diferencias pueden medirse por perturbación de un átomo entre niveles próximos de energía correspondientes a diversos estados estacionarios del mismo. Obtiene así una nueva relación de incertidumbre entre "la precisión E_1 de esa medida de energía" y "el intervalo temporal t_1 en el que el átomo está bajo el influjo de la acción perturbadora", relación que formula y comenta del modo siguiente:

$$E_1\, t_1 \sim h \qquad (3)$$

Esta ecuación corresponde a la ecuación (2) y muestra cómo una determinación precisa de la energía sólo puede obtenerse a costa de una imprecisión correlativa en el tiempo.

Como resumen de ese análisis de los conceptos cuánticos y como generalización de las relaciones de incertidumbre (2) y (3), concluye Heisenberg, en la parte segunda de su artículo: *"Todos los conceptos que son empleados en la teoría clásica para la descripción de un sistema mecánico, pueden definirse también con precisión para los procesos atómicos, por analogía con los conceptos clásicos.* Pero

es un hecho puramente experimental que las experiencias que proporcionan tales definiciones llevan consigo una indeterminación cuando exigimos de ellas la medida simultánea de dos magnitudes canónicas conjugadas. El grado de esta indeterminación está dado por la relación (2), generalizada para cualesquiera magnitudes canónicas conjugadas [como por ejemplo en (3)]".

Y como contrastación epistemológica en que se funda el limitado cambio conceptual de la teoría cuántica, lanza Heisenberg un reto (que recogerá con insistencia Albert Einstein) sobre la generalidad experimental de la imprecisión inherente a la medida cuántica de magnitudes canónicas conjugadas: "Todas las experiencias que podemos emplear para definir esos términos [de posición y velocidad de un electrón], contienen necesariamente la imprecisión dada por la ecuación (2), aun cuando permiten definir exactamente cada uno de los conceptos p y q. Si existiera una experiencia que permitiera una medida simultánea de p y q con precisión mayor que la que corresponde a la ecuación (2), no podría existir la mecánica cuántica. Esta imprecisión que impone la ecuación (2) deja juego para que puedan satisfacerse las relaciones que encuentran su forma más expresiva en la relación de conmutación (1). Esa imprecisión permite que se satisfaga esta ecuación, sin que sea necesario cambiar el significado físico de las magnitudes p y q".

Heisenberg es consciente de las consecuencias filosóficas que encierran esas relaciones de incertidumbre, que caracterizarán a la nueva física cuántica. La física newtoniana había permanecido siempre esencialmente determinista. Pierre Simon Laplace, en 1814, había expresado vigorosamente sus convicciones del determinismo total en el curso del universo y de que toda nuestra incertidumbre se debe a la limitación de nuestro conocimiento sobre él: "Una inteligencia que conociese en un instante dado todas las fuerzas de la naturaleza viva... podría abarcar en una misma fórmula los movimientos de los cuerpos más grandes del universo y los de los átomos más ligeros; nada sería incierto para ella, y el futuro como también el pasado estarían presentes ante sus ojos... Todos los esfuerzos por investigar la verdad aspiran a acercar el espíritu humano a esa inteligencia, de la que sin embargo permanecerá siempre inmensamente alejado... El cálculo de probabilidades se apoya en parte en esa incertidumbre y en parte en nuestros conocimientos..."

Ese sueño cientista se ha esfumado para siempre. Heisenberg describe como conclusión de su artículo cómo ese determinismo físico, o como ha venido en llamarse esa "ley causal" (aplicación física muy específica del principio metafísico de causalidad), resulta invalidado por las relaciones de incertidumbre: "En la formulación exacta de la ley causal, si conocemos con precisión el presente, podemos calcular el futuro, no resulta falsa la consecuencia, sino la premisa. *No podemos* por principio conocer en todos sus detalles el presente. Por tanto, toda percepción encierra una elección entre una multitud de posibilidades y una limitación de la posibilidad futura".

Un último problema es el carácter onto-lógico de esa incertidumbre y esa invalidación de la "ley causal". ¿Se trata de un puro defecto lógico de nuestro conocimiento del mundo físico, según pensaba Laplace, o es un verdadero determinismo óntico, intrínseco a la realidad del mundo físico en que vivimos? Heisenberg, apoyándose de nuevo en su discreto positivismo físico, hace profesión de la armonía onto-lógica de su concepción del mundo y de nuestro conocimiento sobre él. "Podría llegarse a sospechar que bajo el mundo estadístico que percibimos se esconde todavía un mundo verdadero, en el que la ley causal es válida. Pero tales especulaciones, lo subrayamos expresamente, nos parecen infructuosas y sin sentido. La física debe describir formalmente las relaciones entre las percepciones. La verdadera situación puede más bien caracterizarse así: ya que toda experiencia está sometida a las leyes de la mecánica cuántica, y, por tanto, a la ecuación (2), la mecánica cuántica establece definitivamente la invalidez de la ley causal."

Así concluye el famoso artículo de Heisenberg, cuyo cincuentenario celebramos. En una nota añadida al corregir pruebas se insinúan ya las ideas ulteriores del otro pensador de la escuela de Copenhague, Niels Bohr, "que aparecerán pronto en un trabajo sobre la estructura conceptual de la teoría cuántica". Ellas se encargarán de imponer en toda su hondura epistemológica esta revolución conceptual de la física cuántica.

Desarme genético

Gran parte de los "escenarios hipotéticos para el peor caso posible", de los que se postulan para ilustrar la preocupación de algunos observadores acerca de los riesgos potenciales de las investigaciones referentes a la propagación de moléculas de ADN recombinante en huéspedes bacterianos, se centran en la posibilidad de que los experimentadores, al trabajar con la nueva técnica usada para enlazar genes, pudiesen crear por descuido una nueva forma patógena de la bacteria huésped. Y ésta fuera tal, que accidentalmente pudiera escaparse una cepa "maleante" del laboratorio, que pudiese arreglárselas para sobrevivir y multiplicarse en el medio ambiente natural (o hallar una manera de transferir su patogenicidad a otro organismo), que el organismo patógeno recién creado pudiera demostrar que ha adquirido en algún momento y de algún modo una extraordinaria resistencia ante los procedimientos ordinarios disponibles para su erradicación, y que pudiera, por tanto, conducir a una epidemia incontrolable entre los seres humanos u otros organismos.

A tales secuencias hipotéticas de sucesos, a las que la mayoría de los epidemiólogos y demás expertos en enfermedades infecciosas consideran sumamente improbables, se las ha caracterizado en general como montajes de ciencia ficción. Pese a la remotísima probabilidad de que alguna vez se hagan realidad esas conjeturas, han resultado ser difíciles de descartar del todo, debido en parte a ciertas preguntas no contestadas sobre la biología y la genética de la célula huésped que se ha adoptado generalmente para dichas investigaciones: una cepa de la bacteria intestinal omnipresente, *Escherichia coli.*

A causa de estas incertidumbres –y de la ansiedad que era razonable esperar que engendraran en el público–, los investigadores dedicados a las primeras fases de este campo de experimentación, reunidos en la "Conferencia internacional sobre moléculas de ADN recombinante", celebrada en Asilomar, California, en febrero de 1975, convinieron en que la mayoría de los trabajos en marcha entonces (pero no todos) debían proseguirse sólo si "se emplean las justas medidas protectoras, principalmente barreras biológicas y físicas adecuadas para tener a raya a los organismos de nueva creación". Los procedimientos específicos de contención recomendados por los participantes en la conferencia de Asilomar constituyeron el esquema de un conjunto más detallado y algo más estricto de directrices emitidas el pasado mes de junio por los National Institutes of Health (NIH), de los Estados Unidos, para regir la conducta de todas las investigaciones sobre moléculas de ADN recombinante que cuenten con su patrocinio. Se espera, a su vez, que las directrices de NIH, actualmente en vigor para todas las investigaciones sobre este tema que reciban fondos federales, pasen a integrar la legislación federal de los Estados Unidos, extendiéndose así la exigencia de una doble barrera a todas

las investigaciones sobre ADN recombinante.

En su forma actual, las directrices de los NIH exigen 4 niveles de barrera (contención) física, designados *P*1, *P*2, *P*3 y *P*4, correspondientes a diversos experimentos en perspectiva que representan 4 niveles de riesgo. Además, los NIH establecieron normas para la contención biológica, que es un nuevo concepto de seguridad en el laboratorio, basado principalmente (según lo expresó el comité organizador de la conferencia de Asilomar) en el hecho de "que sólo puedan construirse genéticamente bacterias y portadores especiales que posean una capacidad restringida para multiplicarse fuera del laboratorio, y que, además, el empleo de tales organismos refuerce, en muchos órdenes de magnitud, la seguridad de los experimentos con ADN recombinante". Las directrices de los NIH definen los criterios especiales respecto de la contención biológica en función de tres clases diferentes de sistemas de huésped-portador, designados *EK*1, *EK*2 y *EK*3, que cuentan con decrecientes probabilidades de sobrevivir fuera del laboratorio.

El sistema *EK*1 de huésped-portador es simplemente la cepa corriente de *E. coli* adaptada para laboratorio (designada *K*-12), usada conjuntamente con uno de dos portadores que forman clones de ADN recombinante: el virus bacteriano conocido como bacteriófago lambda, o ciertos plásmidos no conjugativos, pequeños bucles de ADN extracromosómico a los que se considera incapaces de transferir genes entre células mediante el proceso conocido como conjugación.

La cepa *K*-12 de *E. coli* se ha cultivado exclusivamente en laboratorio desde 1922, cuando se la aisló de un paciente humano en la facultad de medicina de la Universidad de Stanford. Dicha cepa se ha usado ampliamente en experimentos de genética y de biología molecular desde los años cuarenta, de ahí que se conozcan sus propiedades biológicas y genéticas mejor que las de cualquier otro organismo. A lo largo de su prolongada permanencia en el laboratorio, la cepa *K*-12 ha sufrido numerosas mutaciones, desarrollando una "semitoxicomanía" respecto de las condiciones y nutrientes especiales de su ambiente artificial. Por lo mismo, la *K*-12 ha llegado a ser muy distinta de las docenas de cepas salvajes de *E. coli* que normalmente habitan en el intestino de los animales de sangre caliente. Por ejemplo, en experimentos en los cuales se ha hecho digerir *K*-12 a animales sanos y bien alimentados, incluso a los seres humanos, no se ha podido demostrar que la cepa de laboratorio haya logrado colonizar el tracto intestinal en competencia con sus parientes salvajes más robustos.

Dadas las propiedades especiales de la cepa *K*-12 de *E. coli* y de sus portadores de clonificación prescritos, se ha juzgado que estos sistemas proveen una adecuada medida de contención biológica para la experimentación con ADN recombinante en los casos considerados inocuos o poco peligrosos en potencia. (La designación *EK*1 dada a tales sistemas de huésped-portador deriva de la *E* de *Escherichia* y de la *K* de la *K*-12.)

Para experimentos en los que se estima que el riesgo potencial es mayor, las directrices de los NIH estipulan niveles de contención biológica más elevados que los que proporcionan los sistemas de huésped-portador *EK*1. Así pues, un sistema de huésped-portador *EK*2 es aquel en el cual una amplia variedad de pruebas de laboratorio demuestra que las modificaciones genéticas del huésped y del portador, o de uno de ellos, circunscribe a una en 100 millones la probabilidad de supervivencia de una molécula de ADN recombinante que se hubiese escapado hacia el mundo exterior.

Los trabajos sobre el desarrollo de sistemas adecuados de huésped-portador *EK*2 comenzaron casi inmediatamente después de la reunión de Asilomar a principios de 1975. Hacia enero de 1976, un grupo de experimentadores dirigido por Roy Curtiss III, de la Universidad de Alabama, completó la construcción de una nueva cepa de *E. coli*, derivada de la *K*-12, que incorporaba quince defectos genéticos diferentes, ideados para convertirla en una célula huésped más segura y útil para la experimentación con ADN recombinante. Según Curtiss, la estrategia de su equipo tenía tres aspectos: "Primero, introdujimos una constelación de mutaciones que traía por resultado la destrucción de la célula bacteriana y de su información genética en el caso de que intentara crecer fuera de su medio ambiente *in vitro*, cuidadosamente controlado. Segundo, introdujimos otra constelación de mutaciones que hicieron que la cepa fuese extraordinariamente sensible a diversas sustancias que se encuentran en el medio ambiente o a diversos ambientes físicos, que ocasionarían la muerte de la célula con independencia de los intentos de crecer de dicho organismo. Tercero, introdujimos una constelación de mutaciones que disminuía la probabilidad de transmisión de ADN recombinante a otros microorganismos que pudieran encontrarse en la naturaleza".

La nueva cepa desarmada de *E. coli*, fabricada por Curtiss y sus colaboradores, fue denominada χ1776, "para celebrar el bicentenario de los Estados Unidos." En noviembre pasado, Donald S. Fredrickson, director de los NIH, certificó que la cepa χ1776 satisfacía los requisitos como huésped-portador *EK*2 cuando se usaba conjuntamente con un grupo de portadores clonificantes que eran plásmidos no conjugativos perfeccionados, elaborados por Herbert W. Boyer, de la Universidad de California en San Francisco, y Donald R. Helinski, de la Universidad de California en San Diego.

Más adelante, Frederick R. Blattner, de la Universidad de Wisconsin, Philip Leder, de los NIH, y Philip Sharp, del Instituto de Tecnología de Massachusetts, y los colegas de éstos, diseñaron y construyeron independientemente unos bacteriófagos lambda más seguros; el grupo de Curtiss ha creado otra cepa huésped de *E. coli* (designada χ1953) que puede usarse con varios de estos portadores lambda. Los NIH ya han aprobado algunos de estos sistemas más recientes de huésped-portador en el sentido de que satisfacen las normas *EK*2, y está pendiente la aprobación de otros.

Ante un subcomité parlamentario, y a propósito de la eficacia de la contención biológica en la investigación sobre ADN recombinante, Curtiss sostuvo, en abril del mismo año: "La cepa χ1776 y varios de los demás componentes de los sistemas *EK*2 de huésped-portador han sido reputados como insuficientes por algunos miembros de la comunidad científica... Debe señalarse que el desarrollo de sistemas de contención biológica está en su infancia y que continuamente se llevan a cabo mejoras conforme se obtiene nueva información. En cuanto a las pruebas de seguridad no realizadas aún, forman parte de las pruebas *EK*3, que están comenzando".

Aunque los resultados de las pruebas *EK*3, actualmente en marcha, quizá sirvan para disipar algunas de las preocupaciones encarnadas en la actual proliferación de montajes teatrales de ciencia ficción, es improbable que satisfagan a los críticos de la investigación en el ADN recombinante que consideran dichos experimentos como el "filo de la cuña" que conduce al mal uso deliberado de las técnicas del ADN recombinante para fines torcidos, principalmente en el campo de la bioingeniería humana. Cuesta imaginar qué pruebas de carácter científico pueden aducirse para influir sobre dichos temores, los cuales (para seguir hablando de ciencia ficción) podrían agruparse bajo el encabezamiento de escenarios del tipo de *Un mundo feliz*.

Inmunología del cáncer

Las células cancerosas son cuerpos "extraños" que escapan a la destrucción por el sistema inmunológico. Se está intentando llegar a comprender tal fenómeno y usar la respuesta inmunológica en el tratamiento de esta importante enfermedad

Lloyd J. Old

En el transcurso de la última década ha habido un gran resurgimiento del interés por afrontar el problema del cáncer desde el punto de vista inmunológico, es decir, del estudio de los mecanismos de defensa del organismo frente a invasores extraños. Tales estudios han partido del convencimiento de que, en la célula cancerosa, existe alguna característica especial que la distingue de las células normales, y que esta diferencia puede ser reconocida por el sistema inmunológico. La investigación de la reacción inmunológica contra las células cancerosas y de la paradójica capacidad de muchos tumores para persistir, a pesar de dicha reacción, constituye actualmente un campo de investigación fecundo.

Mi propósito es exponer algunos de los principios de la inmunología del cáncer, tal como los conocemos en este momento, ilustrados con ejemplos ampliamente aceptados, describir algunas líneas de trabajo actuales y plantear algunos interrogantes. Teniendo en cuenta que el objeto de los estudios más extensos y precisos en este terreno ha sido el ratón, mi exposición se refiere sobre todo a los resultados obtenidos con este animal.

Antígenos de la superficie celular

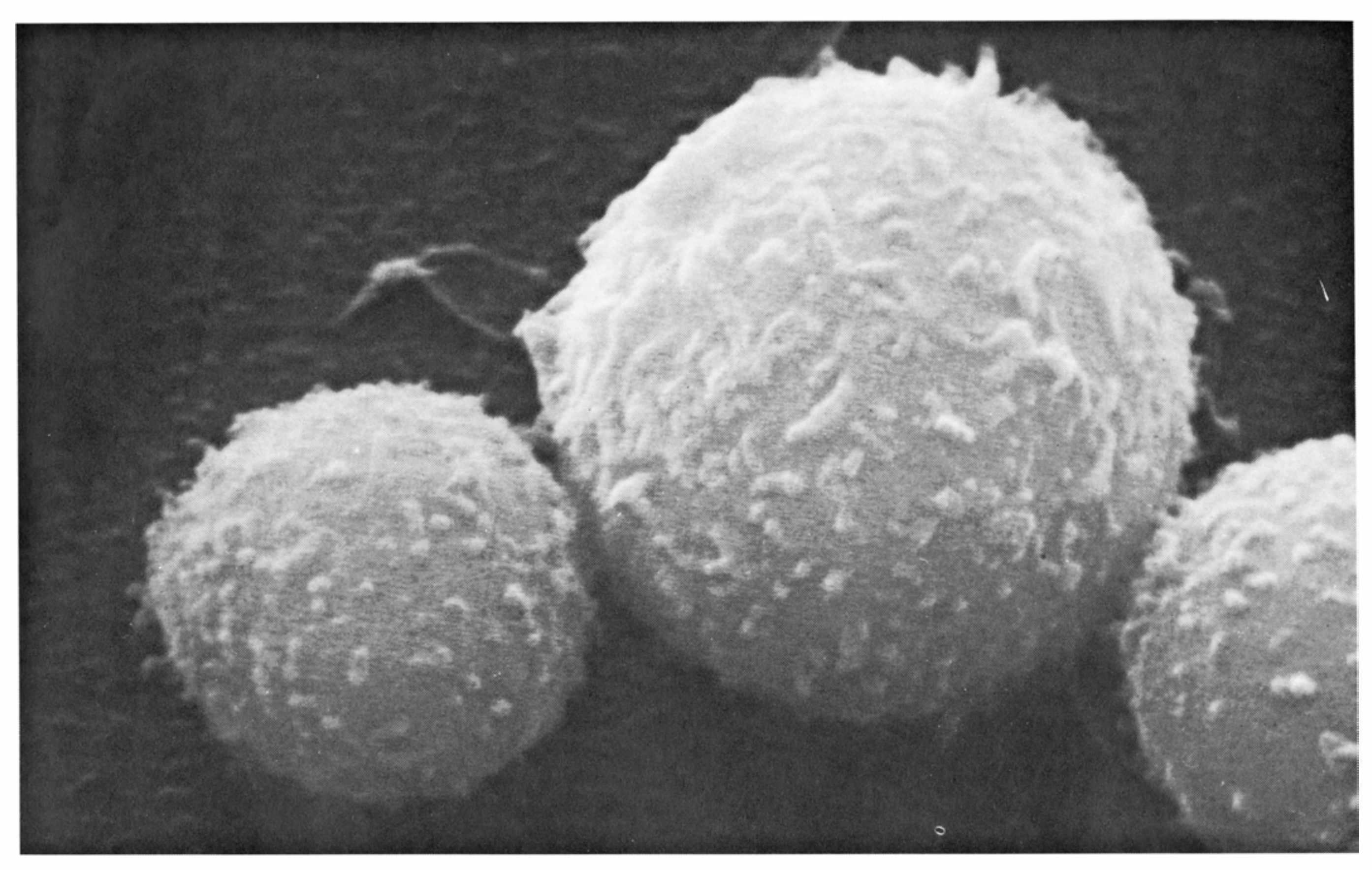
De todas las partes de la célula que pueden alterarse cuando se transforma en cancerosa, la más susceptible es la superficie. Se ha visto que muchos de los mensajes reguladores que controlan el crecimiento y la multiplicación de las células normales tienen su sitio primario de acción en la superficie celular. Es fácil comprender que pequeñas variaciones de la superficie de la célula pueden afectar profundamente el comportamiento de la misma. Se sabe que, cuando se colocan células normales sobre una superficie de vidrio, dejan de migrar y crecer al entrar en contacto unas con otras, fenómeno conocido con el nombre de inhibición por contacto. Sin embargo, si se coloca sobre un vidrio un número equivalente de células cancerosas, éstas, al encontrarse, no dejan de moverse y de crecer sino que siguen proliferando hasta formar una masa celular de múltiples capas. Parece posible, por lo tanto, que una aberración en la estructura normal de la superficie celular pueda explicar la pérdida de la inhibición por contacto, y, como consecuencia, el crecimiento incontrolado de las células malignas.

Esta idea se apoya en la constatación de que las células normales y malignas poseen un repertorio complejo de marcadores moleculares en su superficie externa, compuestos principalmente de proteína y una pequeña cantidad de carbohidrato, y que los marcadores difieren de un tipo celular a otro. Además, cuando se trasplantan células de un animal a otro animal no emparentado, pero de la misma especie, los marcadores de superficie inducen una reacción inmunológica cuya consecuencia es un ataque por parte de los linfocitos del huésped, o sea, por las células que desempeñan un papel crucial en las defensas del organismo contra la infección. Por este motivo estos marcadores moleculares reciben el nombre de antígenos, aunque debe tenerse en cuenta que el hecho de poder inducir una respuesta inmunológica en un animal receptor no emparentado tiene, probablemente, poca relación con la función biológica primaria de la superficie de la célula. En el contexto de la inmunología del cáncer, el descubrimiento de antígenos de superficie, que distinguen entre células cancerosas y normales, ha sido el punto de partida de los intentos de explicar la transformación maligna en función de cambios de la superficie celular y de los intentos de controlar el cáncer por medios inmunológicos.

La demostración inicial de antígenos específicos del cáncer surgió de observaciones llevadas a cabo en ratones de endogamia cerrada. En estas cepas de ratones, como consecuencia de sucesivos cruces entre hermanos, en el curso de muchas generaciones, cada ratón es genéticamente (y, por tanto, antigenéticamente) una réplica de los demás. Los injertos de tejido normal de un animal de una de estas cepas isogénicas suelen ser aceptados por otros ratones de la misma cepa, mientras que son rechazados por animales de otras cepas. Esto significa que los antígenos de las células injertadas son idénticos a los de las células de individuos genéticamente iguales, pero difieren de los antígenos de las células de individuos no emparentados. El descubrimiento de estas "leyes" de la inmunología del trasplante contribuyó a explicar el antiguo misterio de por qué casi en todos los casos, los injertos de órganos o de piel entre personas, intentados muchas veces en la medicina premoderna, no sobrevivían más allá de un breve plazo: los antígenos de la superficie de las células injertadas inducían una respuesta inmunológica en el receptor, carente de los mismos, ocasionando la destrucción del tejido injertado. Se supone que los denominados antígenos de trasplante están presentes en todas las células del organismo. En el hombre los antígenos de trasplante más potentes y mejor estudiados son los antígenos *H-LA* y en el ratón los antígenos *H–2*.

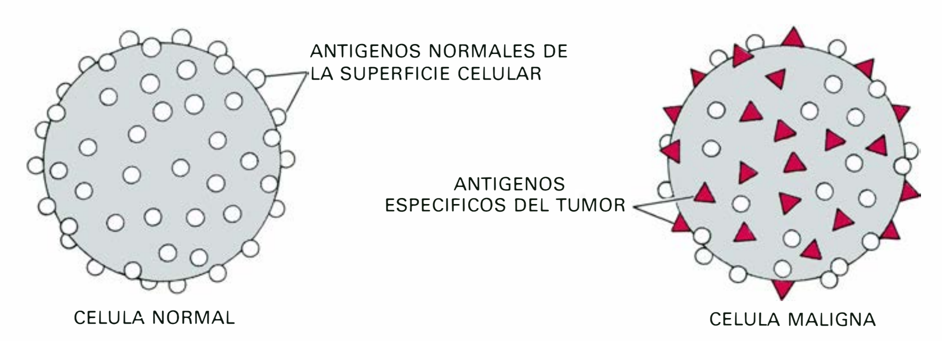
Un mal comienzo

Cuando, a principios de siglo, se inició el estudio de la inmunología tumo-



UN LINFOCITO ATACA a una célula cancerosa. (Microfotografía de microscopio de barrido electrónico, hecha por Andrejs Liepins, del Sloan-Kettering Institute for Cancer Research.) La célula cancerosa *(el esferoide mayor a la derecha)*, tiene en su superficie marcadores moleculares (antígenos) que permiten al linfocito sensibilizado *(célula más pequeña, a la izquierda)* atacarla de manera selectiva y matarla.

LA MUERTE DE UNA CELULA CANCEROSA viene indicada por las ampollas, o profundos pliegues, que aparecen en su membrana. Los linfocitos destruyen las células tumorales, muy probablemente, mediante la liberación de un factor tóxico que rompe la membrana celular.



LOS ANTIGENOS TUMORALES ESPECIFICOS aparecen como consecuencia de la transformación maligna, en la superficie de la mayoría, o quizá de todas las células tumorales. Al igual que los antígenos de trasplante, que existen en las células normales y en las cancerosas, se supone que los antígenos tumorales específicos son unos complejos de proteína e hidrato de carbono sintetizados dentro de la célula e insertos en la membrana celular. Puesto que los antígenos tumorales se diferencian de los antígenos nativos de la célula, deberían desencadenar una respuesta inmune.

ral, todavía no se habían obtenido cepas isogénicas de ratones. Los primeros investigadores encontraron que cuando se trasplantaban tumores aparecidos en unos ratones a otros ratones, el injerto empezaba a crecer al principio para después encogerse y desaparecer bajo el feroz ataque de los linfocitos del receptor. Fenómenos como el descrito llevaron a la creencia errónea de que las curas representaban reacciones inmunológicas dirigidas específicamente contra antígenos específicos para el cáncer.

El descubrimiento subsiguiente de los antígenos de trasplante, existentes tanto en las células normales como en las cancerosas, hizo añicos estas ilusiones. Pronto se vio que, debido a que los primeros experimentos se habían realizado con animales no isogénicos, las referencias a inmunidad específica antitumoral describían de hecho una inmunidad dirigida más bien contra injertos genéticamente distintos y no contra el tumor como tal. Esto quedó claro cuando se vio que la resistencia a los tumores injertados también se podía inducir por inmunización mediante injertos de tejidos normales. A partir de ese momento, los investigadores se dieron cuenta de que la demostración de inmunidad específica antitumoral requería el estudio de tumores originados en ratones de una misma cepa isogénica que tuvieran en sus células antígenos de trasplante idénticos. Si se pudiera demostrar que un tumor de un individuo de esos se comportaba como un injerto de tejido extraño y fuera rechazado por un ratón genéticamente idéntico, ello constituiría una demostración convincente de la existencia de antígenos tumorales específicos, distintos de los de trasplante.

Tumores inducidos químicamente

El primer paso importante en la investigación de antígenos tumorales específicos se dio en 1943, cuando Ludwik Gross, actualmente en el Veterans Administration Hospital del Bronx, Nueva York, descubrió que se podía inmunizar a ratones isogénicos contra un tumor desarrollado en un ratón de la misma cepa isogénica. Por aquel entonces, el trabajo de Gross no recibió mucha atención, debido al pesimismo que imperaba en relación con los estudios inmunológicos del cáncer y pasaron diez años hasta que Edward J. Foley, de la Schering Corporation, aportó nuevas pruebas sobre la existencia de antígenos tumorales específicos en una serie de tumores del ratón. En sus experimentos, Foley utilizó tumores inducidos con metilcolantreno, un hidrato de carbono policíclico capaz de producir cáncer en ratones, al cabo de dos a cuatro meses después de inyectarlo bajo la piel. Estos tumores se indujeron en ratones isogénicos, de modo que se les podía mantener por trasplante seriado de células tumorales a ratones de la misma cepa isogénica, en los que se multiplicaban hasta la muerte del animal. Foley observó que el crecimiento de un trasplante tumoral, seguido de su eliminación, aumentaba la resistencia del ratón a un segundo trasplante del mismo tumor.

En 1957, Richmond T. Prehn y Joan M. Main, del National Cancer Institute, ampliaron el trabajo de Foley demostrando que un ratón isogénico que había rechazado un tumor procedente de un dador genéticamente idéntico solía aceptar un trasplante de piel del mismo animal en que se había originado el tumor. Los trasplantes de piel normal sobrevivían permanentemente, de modo que el rechazo del tumor se podía considerar como prueba concluyente de la existencia de antígenos específicos del tumor que habían surgido durante la transformación maligna de la célula normal a célula cancerosa.

Sin embargo, ni siquiera estos resultados fueron aceptados universalmente, porque se seguía manteniendo la posibilidad de que hubiera diferencias genéticas residuales, no detectadas, entre los miembros de una misma cepa isogénica. Esta crítica fue rebatida por George y Eva Klein, del Real Instituto Carolino de Estocolmo, al demostrar que el propio animal en el que se había inducido originalmente el tumor podía ser inmunizado contra las células de su propio tumor, eliminando así toda duda respecto de que la resistencia al tumor pudiera deberse únicamente a antígenos de trasplante. Estas conclusiones iniciales fueron confirmadas luego en muchos laboratorios de todo el mundo. Unos estudios realizados con sustancias químicas cancerígenas distintas del metilcolantreno demostraron que éstas también inducían tumores que eran antigénicos en la cepa isogénica originaria.

Característica constante y notable de los tumores inducidos químicamente es que cada tumor induce inmunidad contra sí mismo, pero no contra otros tumores. Es decir, los tumores inducidos por un mismo agente químico en una misma cepa isogénica presentan antígenos únicos. Incluso se puede demostrar que, cuando en un mismo animal se inducen dos tumores, éstos presentan antígenos distintos. El origen genético de estos antígenos no se conoce, aunque ha sido objeto de muchas hipótesis. Las dos probabilidades más seriamente contempladas son las siguientes: la información que produce antígenos radica en genes que han mutado (como resultado directo de la interacción de la sustancia cancerígena con el ADN de la célula o como resultado indirecto de la transformación maligna) o en genes generalmente "silenciosos" en el animal adulto, pero que serían activos en otros momentos, posiblemente en la vida fetal. Existen otras explicaciones, pero hasta que estos antígenos queden caracterizados en términos más precisos que los que pueden proporcionar las técnicas de trasplante, su origen seguirá siendo oscuro.

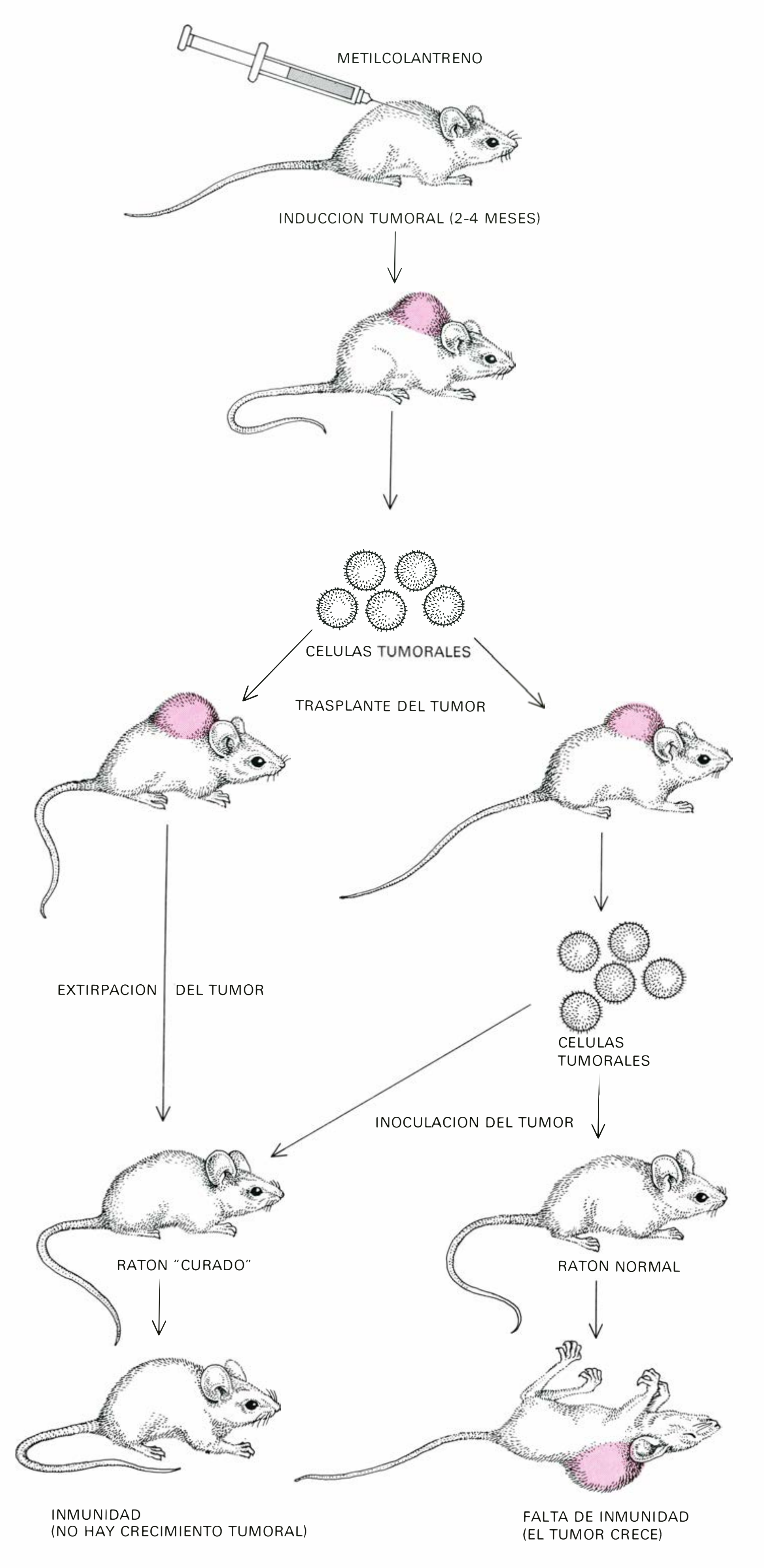
Tumores inducidos por virus

Además de los tumores que se pueden inducir artificialmente mediante sustancias químicas o agentes físicos, como los rayos X, los ratones desarrollan espontáneamente una serie de tumores cuya frecuencia de aparición aumenta con la edad. Cada cepa isogénica de ratones tiene una resistencia o una

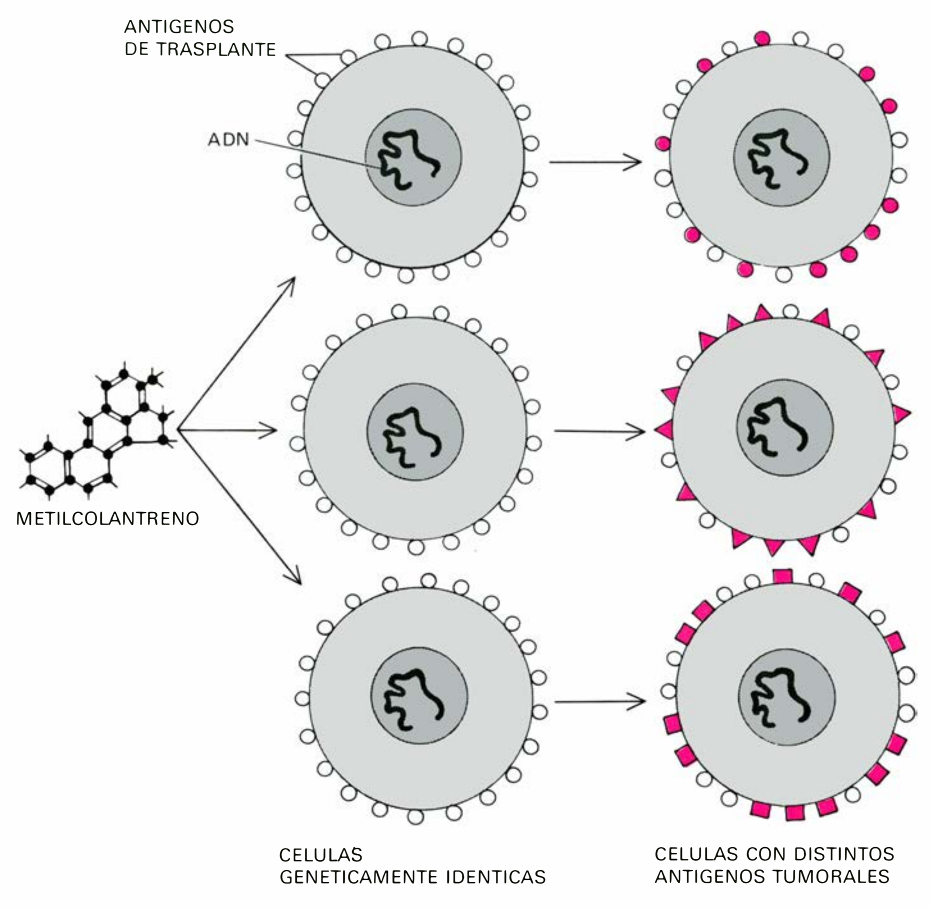
susceptibilidad al desarrollo de tipos específicos de cáncer que son características; algunas tienen una elevada incidencia de leucemias, otras desarrollan cáncer mamario y otras terceras están libres de cáncer. Los tumores que aparecen espontáneamente se pueden dividir en dos categorías. Una categoría abarca los tumores que aparecen sin causa conocida. La otra incluye los tumores inducidos por agentes víricos. Estos agentes son el virus ARN de la leucemia, descrito por Gross, y el virus ARN de tumor mamario, descubierto por John J. Bittner, de la Escuela de Medicina de la Universidad de Minnesota. También se han identificado ciertos virus que si bien pueden ocasionar leucemias u otros tipos de cáncer en los ratones de laboratorio, en circunstancias normales no suelen hacerlo. Sin embargo, han demostrado ser instrumentos de laboratorio insustituibles para dilucidar la forma en que los virus producen el cáncer, y para desenmascarar los factores inmunológicos que influyen en el desarrollo y crecimiento de los tumores.

Uno de estos virus, el virus ADN del polioma, descubierto independientemente por Gross y por Sarah E. Stewart y Bernice E. Eddy, de los National Institutes of Health, es frecuente en ciertas poblaciones de ratones, pero en condiciones naturales los tumores causados por este virus aparecen muy raramente, o nunca, en ratones normales. Los ratones adultos inoculados con virus del polioma no desarrollan tumores. Sin embargo, puede inducirse la aparición de tumores si se inocula el virus a ratones jóvenes (con un sistema inmunológico débil), o a ratones adultos tarados inmunológicamente por procedimientos tales como la extirpación del timo en el momento de nacer o la exposición de todo el animal al efecto de los rayos X. Los tumores de polioma que aparecen en estas circunstancias pueden trasplantarse con éxito a ratones adultos normales de la misma cepa isogénica.

Karl Habel, de los National Institutes of Health, y Hans O. Sjögren, y su colaborador George Klein, ambos en el Real



LA INMUNIDAD a tumores inducidos por agentes químicos se observa en los ratones. Los tumores aparecen de dos a cuatro meses después de la inyección de metilcolantreno bajo la piel. Las células tumorales pueden trasplantarse luego a otros ratones de la misma cepa isogénica y crear tumores en ellos. Si a un ratón receptor se le extirpa el tumor se vuelve resistente a futuras inyecciones del mismo tumor, pero si se inyectan células tumorales a otro normal de tal cepa, éstas proliferan y lo matan.



ANTIGENOS SINGULARES, que aparecen en cada tumor inducido en ratones por la inyección de hidrocarburos como el metilcolantreno y el benzopireno. El mecanismo por el cual se genera esta enorme variedad antigénica tal vez implique mutaciones genéticas o la activación de genes que normalmente no se expresan en el adulto, pero que cumplen alguna función en la vida fetal.

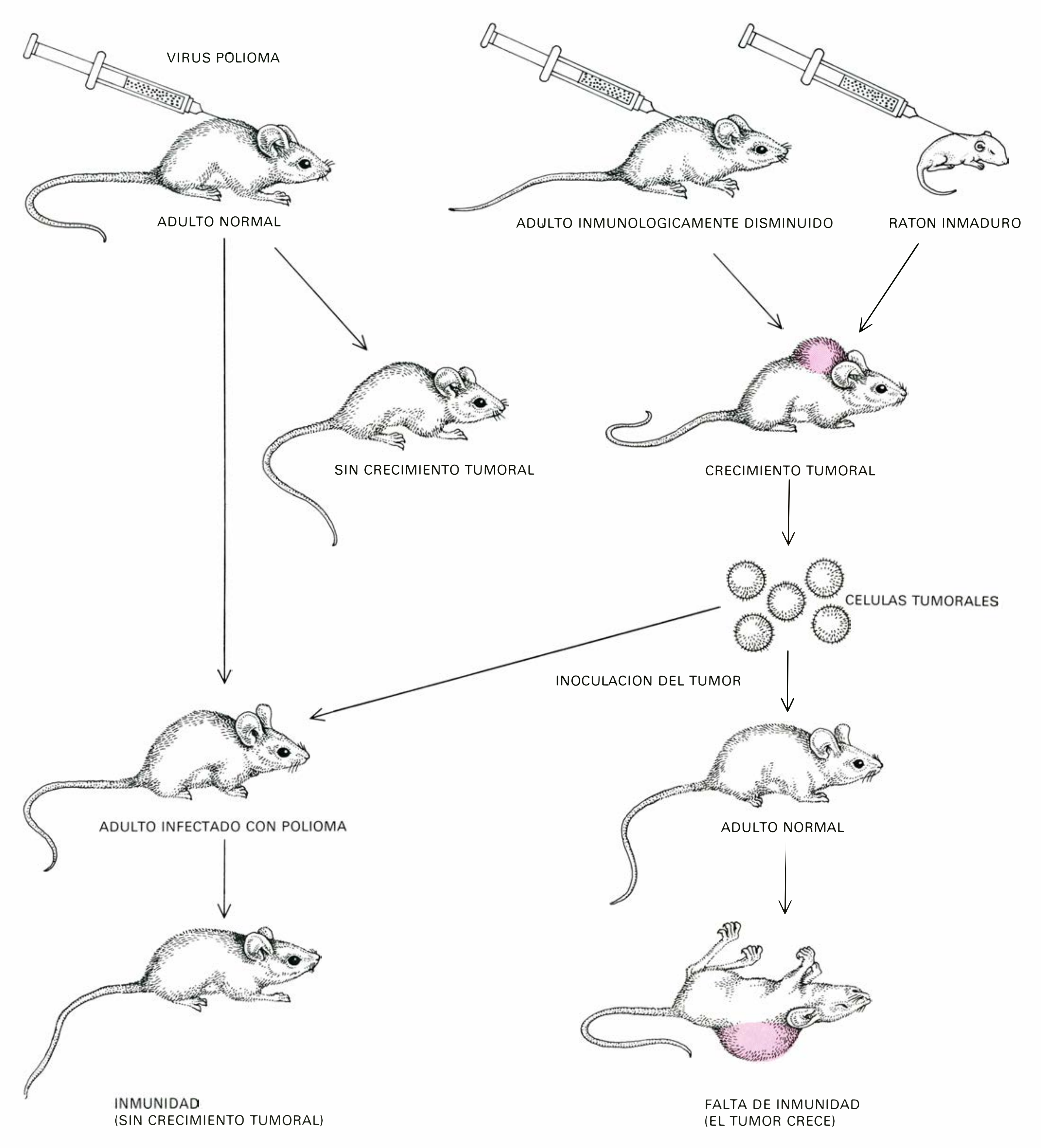
Instituto Carolino, hicieron la observación sorprendente de que estos tumores trasplantados no crecen si el ratón receptor ha sido inoculado, de adulto, con el virus del polioma. La mejor forma de comprender este descubrimiento –de que la inmunización con virus oncogénico confiere resistencia a la inoculación de células tumorales inducidas por el mismo virus– es suponer que, tras la inoculación inicial del ratón adulto con virus polioma, aparecen células tumorales, pero que estas células son reconocidas como extrañas por el sistema inmunológico del animal, y son rechazadas, dejando al animal libre de tumores. Cuando, más tarde, este ratón reciba células tumorales de polioma de otro animal, éstas quedarán rápidamente eliminadas, porque el sistema inmunológico del ratón ha sido previamente alertado por el primer crecimiento y rechazo de sus propias células tumorales. En contraposición con la singularidad antigénica de cada tumor inducido por productos químicos, los estudios de trasplante han demostrado que todos los tumores inducidos por virus polioma tienen el mismo antígeno de superficie celular especificado por el virus, por lo que la inmunización con cualquier tumor de polioma confiere resistencia contra cualquier otro. El estudio de tumores inducidos por otros virus cancerígenos, como el SV40 o adenovirus, ha llevado a la misma conclusión: aparecen antígenos idénticos en distintos tumores inducidos por un mismo virus (aunque cada virus cancerígeno induce la aparición de un antígeno diferente).

Los distintos tumores inducidos por cualquier agente, sea químico o vírico, varían mucho en su capacidad inmunizante. Los antígenos tumorales específicos se describen como "débiles" o "fuertes", según su eficacia para hacer inmunes a los ratones frente a una nueva inoculación con el mismo tumor. Generalmente, se ha encontrado que, en el ratón, los tumores de aparición espontánea son débilmente antigénicos; e, incluso, en algunos casos, parecen no ser antigénicos. En determinadas ocasiones sabemos el motivo por el cual los antígenos no pueden detectarse, lo cual se explicará más adelante. En otras, el método utilizado para detectar el antígeno no es lo suficientemente sensible. No se sabe, por tanto, si los antígenos tumorales específicos son una característica de todos los tumores, sin excepción, o si, por el contrario, las células pueden transformarse en cancerosas sin adquirir estas características distintivas. Sería prematuro formular una opinión sobre este aspecto crucial a partir de la información actualmente disponible.

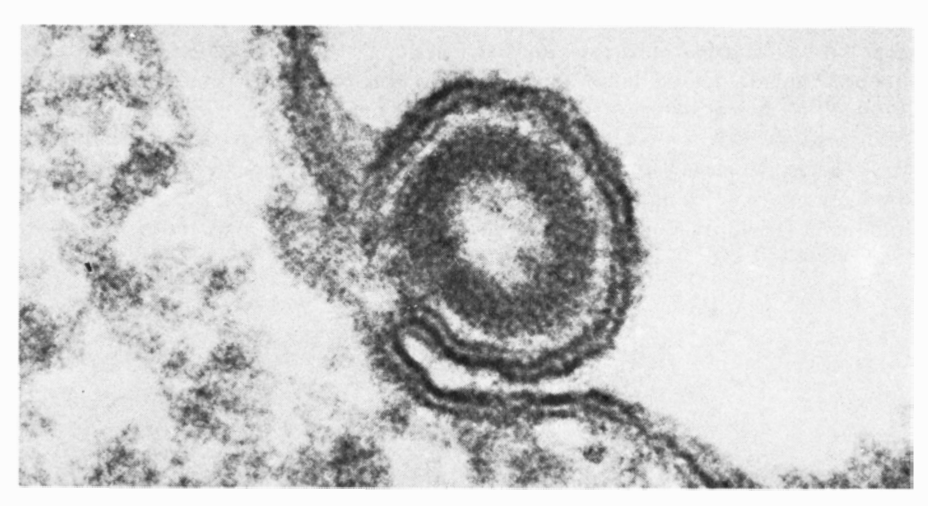
Inmunidad tumoral

Hasta ahora me he limitado a discutir la detección de los antígenos de superficie celular y he hablado muy poco de la naturaleza de la respuesta inmunológica del huésped frente a estos antígenos lo cual es obviamente de primordial importancia si se quiere manipular dicha respuesta ventajosamente para el huésped. El rechazo de tumores, como el rechazo de los tejidos de un dador no emparentado, está mediado primordialmente por los linfocitos. Desde el principio de la década de los 60 se sabe que existen dos clases distintas de linfocitos; ambos se originan de células indiferenciadas primitivas de la médula ósea y migran a distintos órganos en donde se diferencian en células capaces de interaccionar específicamente con los antígenos. Una vía de diferenciación lleva al desarrollo de linfocitos *B* o células *B*. Estas células interaccionan con el antígeno y, a continuación, sintetizan y secretan anticuerpos: proteínas complejas que poseen lados activos, capaces de combinarse específicamente con el antígeno estimulante. La otra vía de diferenciación pasa por el timo, donde las células primitivas se diferencian en distintas clases de linfocitos *T*, o células *T*, responsables de la respuesta inmunológica de tipo celular. Al parecer, las células *T* que han sido expuestas a un antígeno sintetizan en su superficie moléculas análogas a anticuerpos, con las cuales reconocen al antígeno en futuros contactos, iniciando una serie de reacciones que pueden dar lugar a la destrucción de células que tengan antígenos de trasplante extraños o antígenos específicos de tumor en su superficie. Las células *T*, además, segregan factores que median interacciones complejas entre células *T*, células *B* y otras células del sistema inmunológico, especialmente las grandes células necrofágicas llamadas macrófagos. Resumiendo, tanto los aspectos humorales (anticuerpos), como celulares (células *T* y macrófagos) de la respuesta inmunológica parecen estar involucrados en la inmunidad tumoral, aunque quizá la más indispensable sea la inmunidad de las células *T*.

Tras el reconocimiento de los antígenos específicos de tumor, los inmunólogos buscaron métodos analíticos que fueran más rápidos y menos engorrosos que las técnicas de trasplante, con las cuales el punto de referencia es el rechazo del tumor por ratones isogénicos. En los últimos cincuenta años se ha desarrollado un conjunto de poderosas técnicas para demostrar reacciones inmunológicas *in vitro,* no ya *in vivo.* Estas técnicas se han aplicado ampliamente al estudio del cáncer, aunque no siempre con suficiente criterio. En los últimos años se ha dedicado un gran esfuerzo al desarrollo de métodos para medir la reactividad de los componentes celulares del sistema inmunológico, particularmente del linfocito *T.* A pesar de que se ha prosperado mucho en este terreno, hace falta una mayor



EL VIRUS POLIOMA induce tumores cuando se inyecta en ratones recién nacidos, carentes de un sistema inmunológico potente, o en ratones adultos cuyo sistema inmunológico ha quedado debilitado por irradiación o por la extracción del timo al nacer. Los ratones adultos sanos son resistentes a la inducción de tumores por polioma, pero aceptan trasplantes de tumores inducidos por polioma en otros ratones pertenecientes a la misma cepa. La infección con virus polioma antes de tales trasplantes protege al receptor contra la proliferación de las células tumorales, presumiblemente porque su sistema inmunológico ha sido alertado por el crecimiento y rechazo de sus propias células tumorales inducidas por polioma. Esta inmunidad cruzada se produce porque todos los tumores inducidos por polioma expresan el mismo antígeno.



VIRUS DE LA LEUCEMIA DEL RATON que brota de la superficie de una célula. Microfotografía electrónica, obtenida por Etienne de Harven, con un aumento de 310.000 diámetros.

EXPRESION DE GENES DE *MuLV*	CEPA DE RATONES	SUPERFICIE CELULAR DEL TIMOCITO NORMAL
COMPLETA	*AKR* (CEPA CON ALTA INCIDENCIA DE LEUCEMIA)	ANTIGENOS DE DIFERENCIACION CELULAR; VIRUS DE LA LEUCEMIA DEL RATON; *GCSA*; G_{IX}; ANTIGENOS SOLUBLES
PARCIAL	129 (CEPA CON BAJA INCIDENCIA DE LEUCEMIA)	G_{IX}; ANTIGENOS SOLUBLES
NULA	*BALB*/c (CEPA CON BAJA INCIDENCIA DE LEUCEMIA)	ANTIGENOS DE DIFERENCIACION CELULAR

LOS ANTIGENOS DE LA LEUCEMIA VIRICA se encuentran en la superficie de las células leucémicas, en los timocitos normales de ratones de cepas seleccionadas por su elevada frecuencia de leucemia y en células expresamente infectadas con virus de la leucemia del ratón *(MuLV)*. Los dos antígenos principales, *GCSA* y G_{IX}, son componentes proteicos de la partícula del *MuLV*. Cuando forman complejos con hidratos de carbono y se insertan en la membrana celular, pasan a ser antígenos de la superficie celular. El *GCSA* aparece únicamente cuando el virus está siendo sintetizado activamente, pero el G_{IX} puede aparecer también en los timocitos normales de algunas cepas en ausencia de producción vírica. Esta y otras observaciones sugieren que los genes del *MuLV* están incorporados al material genético de los ratones pero se expresan en grado vario.

comprensión de las condiciones necesarias para el funcionamiento de la célula *T* en el tubo de ensayo para poder valorar definitivamente los distintos métodos de detectar la inmunidad celular frente al cáncer.

Métodos serológicos

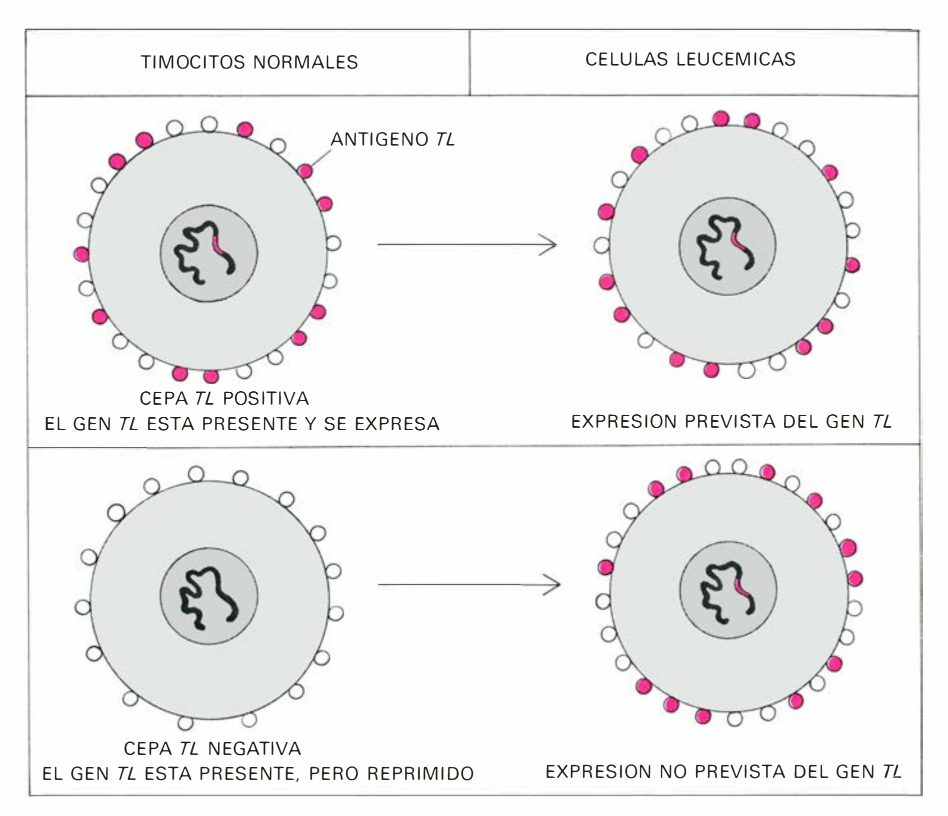
Los métodos más avanzados y versátiles a disposición del inmunólogo son los conocidos colectivamente como métodos serológicos. En ellos se emplea a los anticuerpos como instrumento analítico. Difícilmente se podría exagerar el valor de estos métodos para la investigación y para el laboratorio clínico: dos ejemplos típicos son el diagnóstico de las enfermedades infecciosas y la tipificación ("grupaje") de la sangre para las transfusiones. Se ha dedicado un enorme esfuerzo al intento de detectar antígenos específicos de cáncer mediante la inmunización de caballos, conejos y cabras con tejido canceroso humano y el análisis de los anticuerpos fabricados por el animal, para ver si se puede demostrar que alguno reacciona específicamente con células cancerosas y no reacciona con las células normales. (Gran parte de este trabajo ha sido inspirado por la esperanza, todavía no cumplida, de que un suero específico anticáncer pudiera ser útil en el tratamiento del cáncer en el hombre.) Por regla general, estos esfuerzos han resultado decepcionantes, con pretensiones iniciales de especificidad para el cáncer que, tras un estudio más exhaustivo, han dado lugar a la comprobación, más realista, de que el "antígeno canceroso" también se encuentra en tejidos normales.

De todos modos, aún son grandes las posibilidades de este planteamiento, para lo que puede servir como ilustración el descubrimiento de Garri I. Abelev, de la Academia de Ciencias de la URSS. Abelev inmunizó conejos con sangre de ratones aquejados de un tumor hepático. Después de eliminar del suero del conejo todos los anticuerpos que reaccionaban con la sangre de ratón normal, todavía quedaba, en dicho suero, un anticuerpo que reaccionaba con algún componente de la sangre de ratones cancerosos. Más tarde, el mismo Abelev y otros hicieron observaciones parecidas en humanos. Este antígeno, llamado alfa-fetoproteína, pertenece a una clase de sustancias que son fabricadas durante la vida fetal, para descender a niveles indetectables poco después del nacimiento, probablemente porque los genes que controlan su producción han sido inactivados. Como consecuencia de la transformación maligna, los genes que con-

trolan la producción de estas proteínas fetales pueden reactivarse, y el antígeno puede aparecer de nuevo. La detección de la alfa-fetoproteína en la sangre ha resultado útil para el diagnóstico precoz del cáncer de hígado en humanos. Existen otros antígenos fetales y su detección y análisis se han convertido en un campo activo de trabajo en la inmunología del cáncer.

Una de las pruebas serológicas más útiles para el estudio de los antígenos superficiales de las células cancerosas ha sido el test de citotoxicidad, desarrollado originalmente por Peter Gorer, del Guy's Hospital de Londres. En esta técnica, las células que llevan determinado antígeno en su superficie mueren cuando son incubadas con anticuerpo específico para dicho antígeno, en presencia de un factor del suero sanguíneo, llamado complemento. En realidad, el complemento consiste en una serie de proteínas que, cuando hay anticuerpo unido al antígeno, son activadas secuencialmente en una cascada de reacciones enzimáticas. El complejo activo terminal parece debilitar la membrana de la célula antigénica, llevándola a su lisis o disolución. La secuencia de las reacciones del complemento sirve para ampliar centenares de veces la reacción inicial antígeno-anticuerpo, de modo que la participación de unas pocas moléculas de anticuerpo puede llevar a la destrucción de la célula antigénica. Se ha implicado al complemento en muchos fenómenos inmunológicos, pero su papel más importante en la inmunidad tumoral es el de producir la lisis de las células que portan antígenos tumorales específicos, contra los que se ha producido anticuerpo. En la prueba de citotoxicidad ideada por Gorer, a las células tumorales muertas por anticuerpos se las identifica añadiendo un colorante a la solución; las células lisadas se tiñen con el colorante, pero no así las células intactas. Algunas modificaciones han hecho de este método una prueba muy sensible para detectar y cuantificar antígenos de superficie celular. Como las células leucémicas son más sensibles al anticuerpo citotóxico que otros tipos de células cancerosas, han sido las más intensamente estudiadas.

Leucemia es un término que sirve para designar la proliferación anormal de los leucocitos, tanto en forma de tumores sólidos llamados linfomas, como de células libres en la sangre. En el ratón, este cáncer suele originarse en el timo; ello proporciona al serólogo una magnífica oportunidad para comparar los antígenos de superficie de la célula leucémica con los de la célula tímica normal o timocito. En 1961, Edward A. Boyse, Elisabeth Stockert y yo mismo, en el Sloan-Kettering Institute for Cancer Research, iniciamos nuestro estudio de los antígenos de superficie de células leucémicas y de timocitos normales. En aquel momento solamente se podía demostrar un antígeno en las células linfoides; éste era el antígeno principal de trasplante *H*-2, que está presente en la mayoría de las células del ratón o en todas ellas. En el curso de estos años, con otros colegas, especialmente Tadao Aoki y Toshitada Takahashi, nosotros hemos definido diez antígenos adicionales; se sabe, pues, más de los antígenos de superficie de estas células que de los de cualquier otro tipo celular en el ratón.

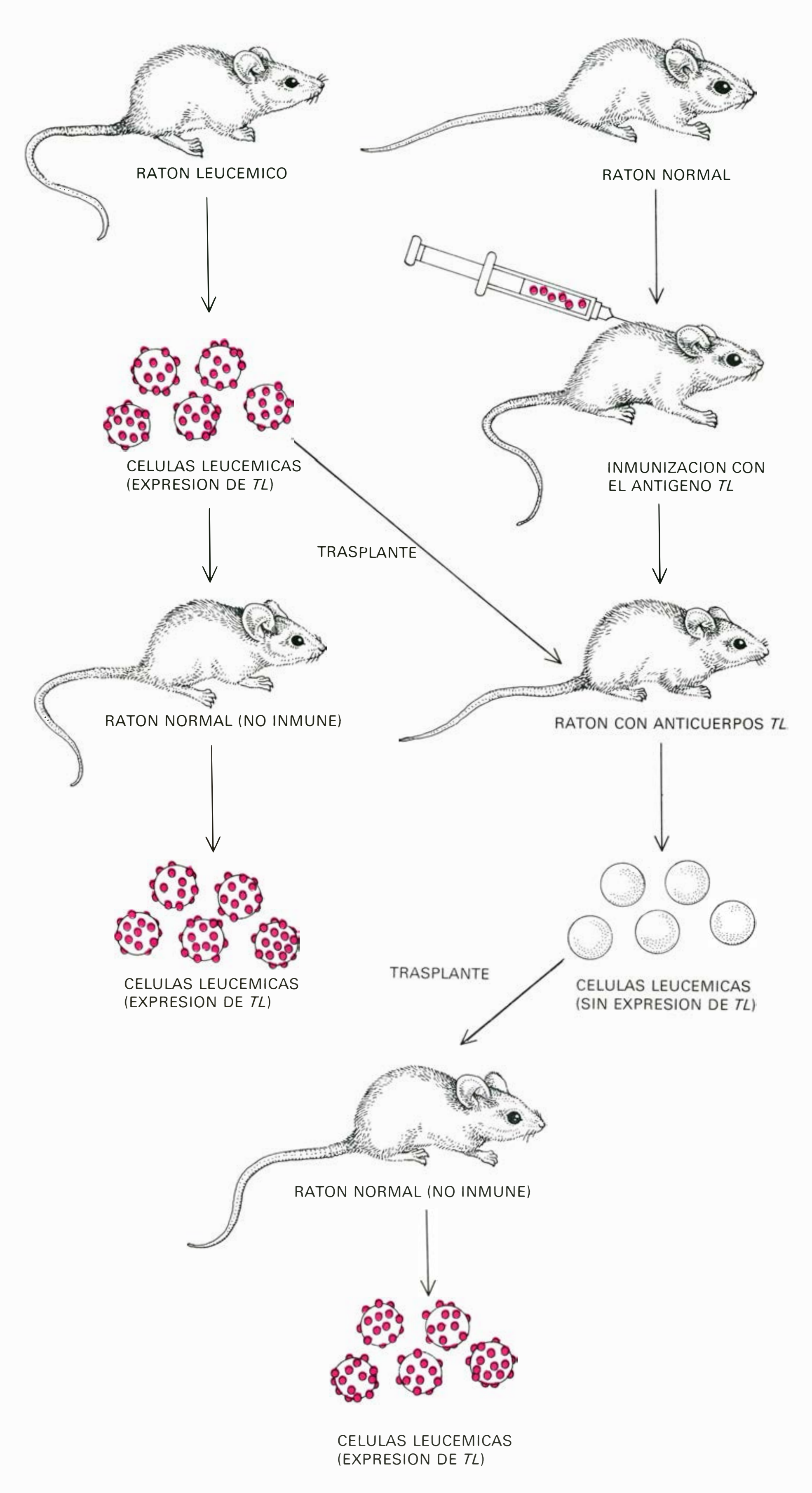


EL ANTIGENO TIMICO-LEUCEMICO *(TL)* **aparece en las células leucémicas y también en los timocitos normales de ciertas cepas isogénicas de ratones, llamadas *TL*-positivas. Otras cepas de ratones carecen de *TL* en sus timocitos normales (cepas *TL*-negativas); cuando desarrollan leucemia, las células leucémicas pueden expresar los antígenos *TL*. Esta observación sugiere que todas las cepas de ratones contienen la información genética para *TL*, pero que en las cepas *TL*-negativas esta información existe en un estado silencioso, que puede activarse durante el proceso que conduce a la transformación de la célula normal en célula maligna.**

Antígenos de leucemia

En los timocitos y células leucémicas se reconocen actualmente, además del *H*-2, otras tres categorías de antígenos de superficie. El primero se puede ilustrar por una serie de antígenos, llamados antígenos *Ly*, que marcan todas las células linfoides derivadas del timo. A estos antígenos se les ha llamado antígenos de diferenciación, porque los genes que controlan su expresión son activados únicamente en células que siguen determinada línea de diferenciación, desde las células embrionarias hasta formar células especializadas. Las células leucémicas de origen tímico llevan antígeno *Ly*, como resultado directo de su tipo celular ancestral. Los genes que determinan los antígenos *Ly* siguen las leyes clásicas de la genética mendeliana. Estos antígenos resultaron de gran interés para los inmunólogos desde que Hiroshi Shiku y Pawel Kisielow y sus colegas descubrieron, en el Sloan-Kettering Institute, que se pueden distinguir tres clases de células *T* con funciones inmunológicas diferentes, según el juego de antígenos *Ly* expresados en su superficie.

La segunda categoría de antígenos de superficie en células leucémicas, y, en ciertas circunstancias, en células normales, son los antígenos víricos relacionados con el virus de la leucemia del ratón, abreviadamente *MuLV* (de "murine leukemia virus"). Gran parte de lo que actualmente se sabe sobre la biología de este virus y de las leucemias que ocasiona se ha conocido a partir del empleo de técnicas serológicas para la detección de antígenos virales. Como veremos, la misma expresión de un antígeno viral puede constituir la única prueba de la presencia de *MuLV*. Se ha identificado una serie de antígenos de *MuLV*, entre los cuales hay componentes estructurales



LA MODULACION ANTIGENICA es un "mecanismo de escape" bien documentado, que permite a las células cancerosas evadir la respuesta inmune del huésped. Cuando se inyectan células leucémicas que contienen *TL* en ratones previamente inmunizados a este antígeno (y que, por tanto, han sintetizado anticuerpo contra el mismo), el *TL* desaparece de la superficie de estas células. Las células leucémicas están funcionalmente "desnudas" y no pueden ser reconocidas por el anticuerpo anti-*TL* o por linfocitos citotóxicos. Cuando se reinyectan estas células moduladas a un ratón no inmunizado, el antígeno *TL* de nuevo vuelve a manifestarse en la superficie celular.

que forman parte de la partícula vírica, antígenos víricos que quedan incorporados en la superficie celular y antígenos víricos solubles que se desprenden de las células infectadas con *MuLV* e ingresan en el medio líquido.

La primera oportunidad de seguir este virus en su huésped natural vino a través de la descripción de lo que se conoce como el antígeno Gross de superficie celular, *GCSA* (de "Gross cell-surface antigen"), el cual se expresa siempre que una célula produce *MuLV*. Encontramos *GCSA* en varios tejidos normales y leucémicos procedentes de cepas de ratones seleccionadas por su gran incidencia de leucemia, como las cepas llamadas *AKR* y *C*58, pero no se detectó en ningún tejido normal procedente de cepas con baja incidencia de leucemia. Lo inesperado fue observar que el *GCSA* podía encontrarse en los tumores sólidos o en las leucemias de dichas cepas con baja incidencia de leucemia. Este fue uno de los primeros indicios de que todos los ratones podían estar infectados por *MuLV* y que otros factores podían determinar si se expresaba o no el virus.

Por el trabajo de Wallace P. Rowe y sus colegas, en los National Institutes of Health, sabemos que, en el ratón, la información genética para el *MuLV* es omnipresente y forma parte integral del genoma (juego completo de genes), y que estos genes víricos se transmiten de generación en generación, junto con los demás genes del ratón. El hecho de que los genes *MuLV* se expresen o no, trayendo como resultado un ratón infectado y la producción de partículas víricas, viene determinado por otros genes que varían de una cepa isogénica a otra. También se han encontrado cepas en las que los genes víricos se expresan en parte produciendo sólo pequeñas cantidades de virus o de proteínas víricas, o que no se expresan en absoluto. Las hormonas, las sustancias cancerígenas o la exposición a los rayos X pueden activar los genes *MuLV* silenciosos, y en algunas cepas de ratones estos agentes pueden incluso causar leucemia.

Sin embargo, la sola expresión completa del *MuLV* no parece ser suficiente para transformar una célula en maligna. Si se calcula el número de proteínas que la información genética contenida en este virus podría codificar, resulta que los productos víricos conocidos corresponden, probablemente, a todos los genes del virus. Puesto que una célula puede expresar todo el repertorio de proteínas víricas sin mostrar ninguna prueba de malignidad, su presencia no es de por sí causa de leucemia. El hecho de que la leucemia se origine generalmente en

el timo, a pesar de que los genes del *MuLV* están probablemente incorporados en el ADN de una de cada dos células nucleadas del ratón, indica que solamente las células que siguen esta línea determinada de diferenciación son sensibles a la acción productora de leucemia de este virus. La forma en que las fuerzas de la diferenciación controlan la expresión y la función de los genes *MuLV* es un interrogante crucial en oncología.

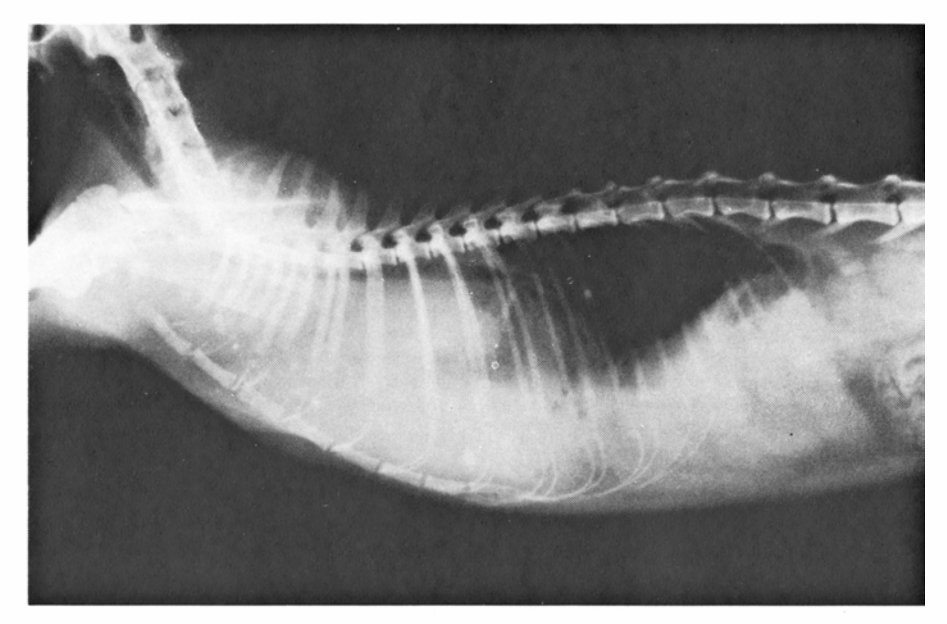
El análisis del G_{IX}, otro antígeno especificado por el *MuLV*, ilustra claramente el efecto de la diferenciación celular en la expresión de los genes víricos. Originalmente el G_{IX} fue reconocido como un antígeno de superficie de los timocitos normales, siendo detectado en unas cepas de ratones y en otras no. Su relación con el *MuLV* se estableció a partir del hecho de que se podía inducir su aparición en células carentes del mismo, mediante la infección deliberada con el virus. Al igual que el *GCSA*, el G_{IX} se expresa siempre que se expresa el *MuLV* y puede aparecer en las células leucémicas de ratones que normalmente no expresan el G_{IX} en sus timocitos. El G_{IX} está asimismo presente en los timocitos normales de ciertas cepas de ratones que no dan muestras de infección con *MuLV*. En esas cepas el G_{IX} se halla sólo en los timocitos y no, en cambio, en ningún otro tejido linfoide.

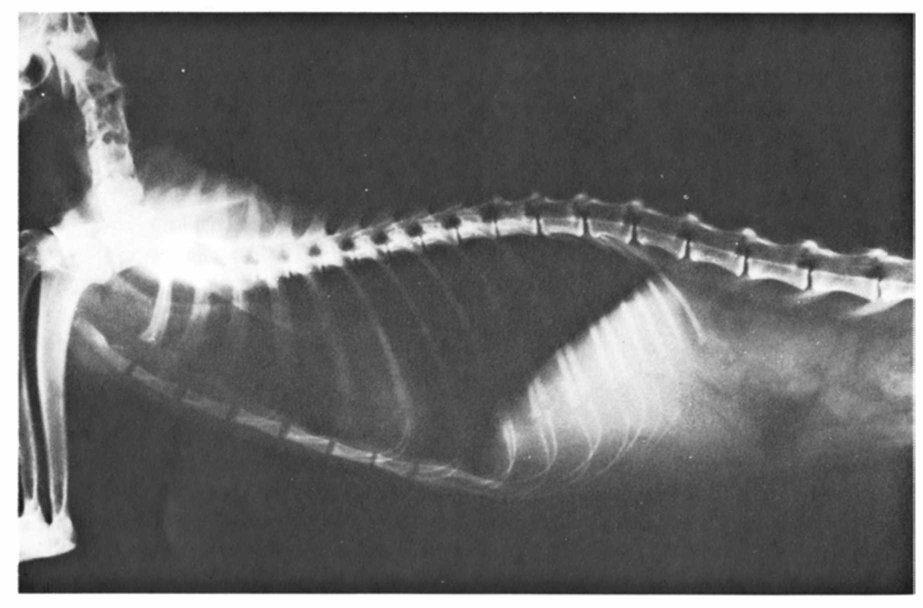
El análisis genético ha revelado otras características únicas del G_{IX}. A diferencia del *H*-2 y otros antígenos corrientes de la superficie celular, que precisan un solo gen para su expresión, parece que el G_{IX} necesita dos genes. Además, mientras otros antígenos de superficie celular son el producto de una u otra forma del mismo gen, las cepas de ratones carentes de G_{IX} no presentan ningún otro producto genético alternativo. Otra peculiaridad del sistema G_{IX} consiste en el hecho de que las cepas de ratones pueden pertenecer a cuatro categorías diferentes, según la cantidad relativa de antígeno G_{IX} que expresan en sus timocitos. Esta característica poco usual puede tener su origen en el distinto número de copias de los genes *MuLV* incorporadas en el material genético de las diversas cepas.

La relación entre estos antígenos leucémicos y el *MuLV* ha quedado más clara por el descubrimiento de que el G_{IX} es un constituyente de la principal glucoproteína de la cubierta de la partícula del *MuLV*, designada *gp*70. Por otra parte, Erwin Fleissner y sus colaboradores, en el Sloan-Kettering Institute, han descubierto recientemente que el *GCSA* está inmunológicamente relacionado con dos de los componentes proteicos internos de la partícula vírica, *p*30 y *p*15.

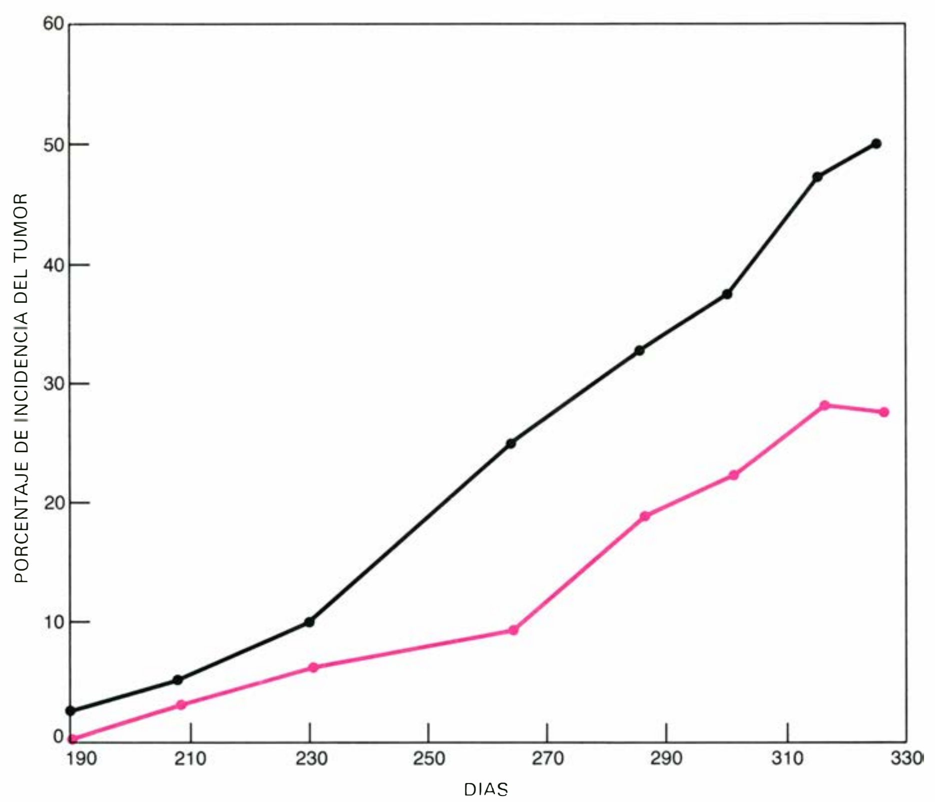
En nuestro análisis del G_{IX} destacan tres importantes puntos. Primero, que la diferenciación celular es de una importancia crucial en la expresión de los genes del *MuLV*. En ratones normales de algunas cepas con baja incidencia de leucemia, el gen G_{IX} sólo se activa en células linfoides que se diferencian en el ambiente tímico. En este sentido el G_{IX} puede considerarse un antígeno de diferenciación comparable a los antígenos *Ly*.

En segundo lugar, la relación entre el *MuLV* y el huésped es íntima, tanto a nivel genético como a nivel de superficie celular. El G_{IX} es un producto de gen vírico que ha sido incorporado como constituyente de la superficie de una célula normal, el timocito. Otra manera de contemplar este descubrimiento consiste en suponer que en el curso de la evolución del ratón los genes de virus y del huésped colaboraron estrechamente para producir la estructura de la superficie de ciertas células. Si no se supiera que el G_{IX} se origina a partir del *MuLV*, al primero se le consideraría como producto de un gen normal de ratón. Por una serie de razones, especialmente la gran frecuencia con que el *MuLV* está presente en los ratones, se sospecha cada vez más que el *MuLV* desempeña una fun-





REGRESION ESPECTACULAR DE UN TUMOR producida en un gato mediante la perfusión de plasma sanguíneo que contenía complemento, conjunto de proteínas que toman parte en la destrucción de células antigénicas en presencia de anticuerpo específico. La radiografía de la parte alta de la ilustración muestra la cavidad torácica del gato enteramente ocupada por un linfoma avanzado *(zona opaca)*. La radiografía de abajo, hecha 12 días después de la perfusión de plasma que contenía complemento, demuestra que ha desaparecido completamente. Es evidente que el déficit de factores inmunológicos secundarios, como el complemento, puede ser responsable de la incapacidad que muestra el sistema inmunológico para destruir ciertos tumores antigénicos.



INHIBICION DE LA APARICION ESPONTANEA de tumores, después de la infección de ratones jóvenes con BCG. Se sabe que los tumores espontáneos de mama de algunas cepas de ratones son causados por el virus Bittner, el cual se transmite de madre a hijo por la leche. Los tumores empiezan a aparecer cuando los ratones llegan a los seis meses de edad y siguen apareciendo gradualmente, de modo que, al año, más de la mitad de los ratones presentan tumores *(línea negra).* **Si se les inyecta BCG a los dos meses, el ritmo de aparición de tumores es lento** *(curva de color).*

ción crítica en la vida normal del huésped, aunque quede por resolver cuál sea.

En tercer lugar, la aparición de G_{IX} y *GCSA* en leucemias y otros tumores en ratones que no expresan estos antígenos durante su vida normal, incide significativamente sobre el origen de los antígenos tumorales específicos. He mencionado la posibilidad de que la activación de genes normalmente silenciosos explique la individualidad singular de los antígenos de tumores inducidos por agentes químicos. El origen genético de la aparición de G_{IX} o *GCSA* en las leucemias de cepas con baja incidencia sería el *MuLV*.

El sistema *TL*

El otro ejemplo de antígeno tumoral específico que debe su origen a la activación genética viene de nuestro descubrimiento, en 1963, del sistema *TL*, que constituye la tercera categoría de antígenos de superficie celular. El *TL* tiene muchas peculiaridades comunes con el G_{IX}, y, aunque no tenemos ninguna prueba al respecto, sospechamos que puede haber algún tipo de asociación con un virus. Este antígeno se llama *TL* porque únicamente se encuentra en timocitos normales de ratón y en células leucémicas. Las cepas normales de ratón se pueden clasificar en *TL* positivas y *TL* negativas, según el grupaje del timo respecto de los antígenos *TL*. La característica poco común del *TL*, que comparte con el G_{IX}, es que cuando, en ratones cuyos timocitos carecen de *TL*, aparecen leucemias espontáneamente, o se las induce mediante agentes físicos o químicos, la célula leucémica puede expresar estos antígenos. Al igual que con el G_{IX}, eso se ha explicado suponiendo que todos los ratones poseen la información genética para *TL;* aunque otros genes controlan el hecho de que se exprese o no en los timocitos normales. La leucemia introduce un cambio en este control genético, dando como resultado la aparición anómala de *TL* en la superficie de la célula. leucémica.

Estos hallazgos plantean una cuestión interesante acerca de los antígenos que los inmunólogos clasifican como tumorales específicos. Tanto el *GCSA* como el G_{IX} o el *TL* pueden considerarse como específicos respecto de tumores, en las condiciones en que aparecen en ratones que nunca los expresan en su vida normal. Sin embargo, estos antígenos no son realmente específicos de tumor en el sentido de que se presenten solamente en células cancerosas, ya que también aparecen como componentes de células normales en otras cepas de ratones. El único antígeno realmente específico de la leucemia que hemos encontrado en el ratón es un componente del sistema *TL*, designado *TL*-4, que no se ha detectado nunca en células normales.

Aunque se ha avanzado bastante, no tenemos más que un conocimiento somero de la enorme variedad de antígenos de superficie de las células normales y malignas del ratón. Se han identificado otros antígenos, pero no se les ha analizado lo suficiente como para situarlos en una u otra categoría. Naturalmente que el fin último de esta investigación no es compilar un inventario de antígenos de superficie celular, sino utilizar estos marcadores antigénicos para llegar a comprender cómo está construida la superficie celular, cómo afecta la cancerización a estas estructuras y cómo responde el sistema inmunológico.

Mecanismos de escape

Los antígenos de trasplante específicos de tumores inducidos por agentes químicos o por virus polioma, y los antígenos G_{IX} y *TL*, que aparecen en las células leucémicas de ratones que carecen de estos antígenos en sus células normales, dan una idea de la cantidad de antígenos, específicos del cáncer, que se ha detectado hasta este momento. En realidad, cuando se usan técnicas suficientemente sensibles, la mayoría de las células cancerosas del ratón muestran poseer antígenos específicos de superficie. ¿Cómo pueden aparecer estos tumores sin provocar una respuesta inmunológica eficaz que suprima su crecimiento?

No existe ninguna explicación satisfactoria para el éxito de los tumores frente a la respuesta inmunológica. Se han propuesto, sin embargo, una serie de mecanismos que permitirían a las células tumorales escapar de las consecuencias de su antigenicidad. Algunos son capciosamente sencillos, tal como el que se ha venido en llamar "escabullirse". El cáncer, al empezar como células individuales, podría no desencadenar el reconocimiento inmunológico en las primeras fases de su crecimiento. Cuando el sistema inmunológico es alertado, el cáncer ya estaría establecido y sería demasiado grande como para que el cuerpo pudiese eliminarlo.

Tal vez el mecanismo mejor estudiado y más sorprendente sea el de la modulación antigénica: la capacidad de las células cancerosas para enmascarar o perder antígeno ante el ataque inmunológico. Fue descubierto por nuestro grupo cuando nos enfrentábamos con la paradoja de que la inmunización de ratones con antígenos *TL* no confería resistencia al crecimiento de células leucémicas con *TL* en su superficie. Estas

células leucémicas crecían en ratones con altos niveles de anticuerpos *TL* tan bien como en ratones no inmunizados, a pesar de que la cantidad de anticuerpo presente hubiese tenido que conferirles buena protección. Boyse, y yo mismo, encontramos que las células leucémicas de los ratones inmunizados habían perdido antígeno *TL*. Al trasplantar estas mismas células a ratones sin anticuerpo, el antígeno reaparecía en su superficie. Evidentemente, no todos los antígenos tumorales sufren la modulación antigénica, porque en este caso no hubiese sido posible detectar, con técnicas de trasplante, los antígenos de los tumores inducidos por agentes químicos o víricos.

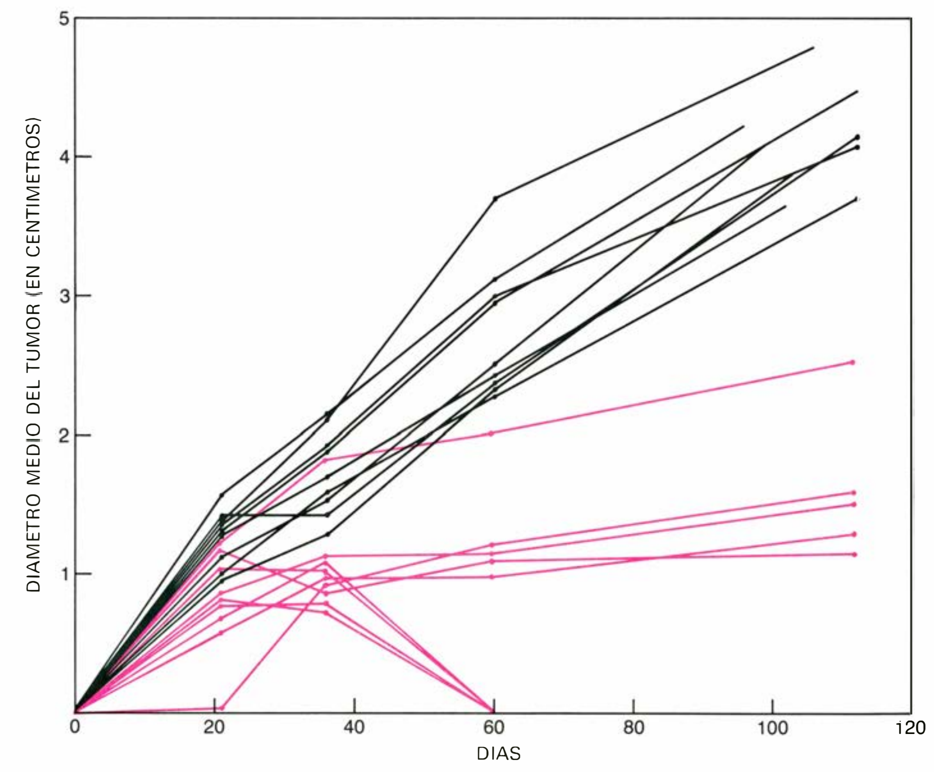
Otra forma que los tumores tienen de escapar a la destrucción inmunológica es inundando el organismo con antígenos tumorales. Los antígenos se unirían, entonces, a los anticuerpos específicos o a los receptores específicos de los linfocitos, impidiéndoles reconocer y destruir las células cancerosas. Llamamos a esta vía de escape "neutralización antigénica".

Los tres mecanismos descritos representan, por parte de la célula tumoral, estrategias que le permiten escapar a la destrucción. Las deficiencias u otras anormalidades, o una reactividad inmunológica específica por parte del huésped, también pueden desempeñar un papel importante en el desarrollo y crecimiento progresivo de tumores antigénicos. Tal como he dicho anteriormente, el virus polioma no puede inducir la aparición de tumores en ratones adultos, a menos que su inmunidad quede reducida artificialmente. Robert A. Good y sus colegas, trabajando entonces en la Escuela de Medicina de la Universidad de Minnesota, descubrieron que las personas con defectos genéticos que afectan a la respuesta inmunológica presentan una incidencia de cáncer mucho más elevada. También es sabido que el cáncer aparece con mayor frecuencia en los pacientes que han sufrido un trasplante y que han de estar sometidos a inmunosupresión para minimizar el rechazo del órgano trasplantado. No obstante, sigue en pie el hecho de que el cáncer puede aparecer en animales de experimentación o en personas cuya respuesta inmunológica general es, a todas luces, perfectamente normal.

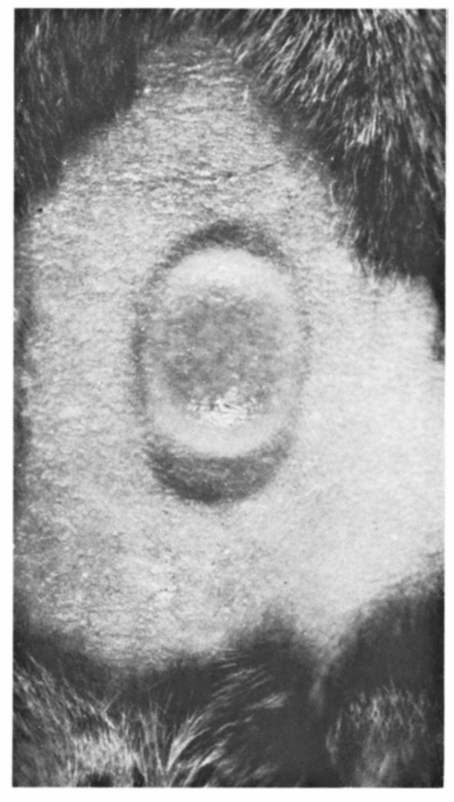
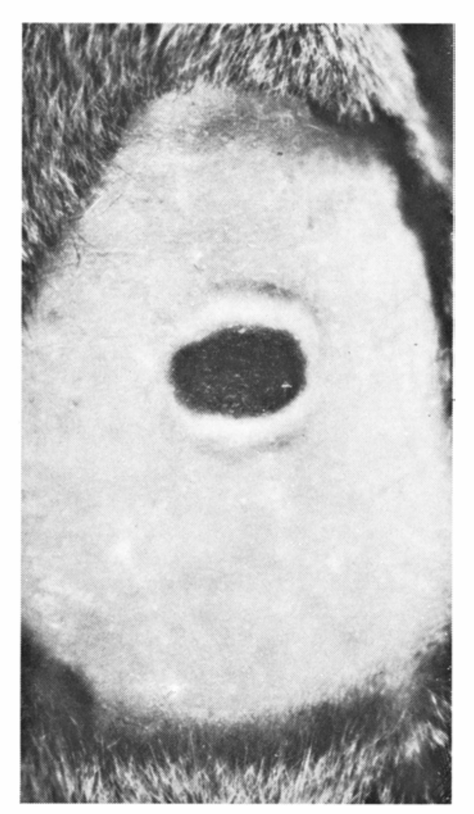
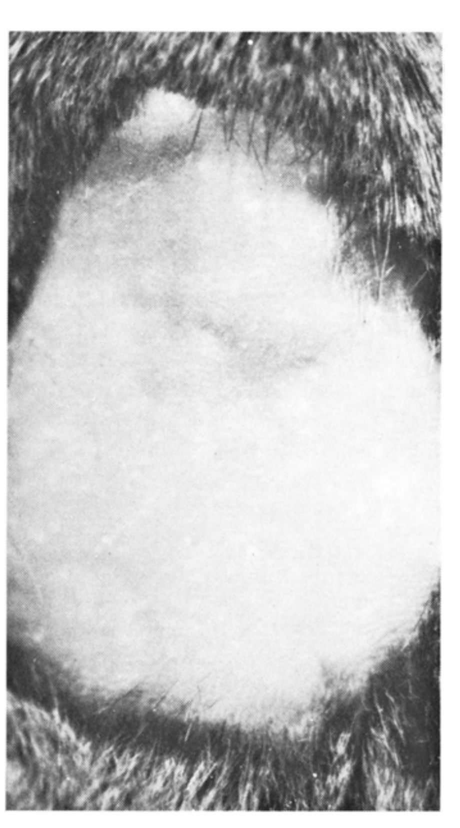
Puesto que actualmente sabemos que las reacciones inmunológicas están sometidas a control genético, la falta de algunos genes *(Ir)* de respuesta inmunológica, necesarios para reconocer determinado antígeno específicamente tumoral, podría permitir la aparición de tumores antigénicos en individuos con un sistema inmunológico normal.

El conocimiento de otro mecanismo para la disminución específica de la inmunidad tumoral en presencia de un sistema inmunológico normal, se debe a los trabajos de Donald L. Morton y de David W. Weiss sobre los tumores de mama en el ratón. La gran mayoría de estos tumores están producidos por el virus Bittner, que se transmite de generación en generación por la leche; los ratones recién nacidos adquieren la infección mamando de madres infectadas. Si los ratones recién nacidos son criados por madres adoptivas no infectadas por el virus, los ratones no se infectan y no desarrollan tumores mamarios. Estudios de trasplante, como los llevados a cabo con los tumores inducidos por agentes químicos, demuestran que los tumores mamarios son mucho más antigénicos en ratones que carecen del virus que en ratones que lo adquirieron en el nacimiento. Aparentemente, los ratones infectados desde el nacimiento llegan a no responder inmunológicamente (se hacen "tolerantes") a ciertos antígenos relacionados con el virus y con el tumor inducido por él y, por tanto, cuando el tumor aparece más tarde en la vida del animal, éste es incapaz de mostrar una respuesta inmunológica.

La mayoría de las explicaciones sobre mecanismos de escape hacen hincapié, en último término, en la ausencia de una inmunidad específica mediada por anticuerpo o por linfocitos. Sin embargo, experimentos realizados por Robert L. Kassel, Elisabeth A. Carswell y por mí, indican que quizás unos factores secundarios necesarios para la acción del anticuerpo limiten de modo crítico ciertas reacciones inmunológicas contra el cáncer. Hemos encontrado extensas destrucciones de células leucémicas después de hacer perfusiones de plasma de ratón sano a ratones leucémicos. El factor activo de este plasma ha resultado ser uno de los componentes del sistema del complemento, el cual, como ya lo he señalado, consiste en una serie de proteínas sanguíneas que participan en una amplia variedad de reacciones inmunológicas, entre ellas la destrucción de células por anticuerpo. La información actualmente disponible indica que, probablemente, algunos animales leucémicos elaboran anticuerpos específicos en cantidad adecuada, pero que, por falta de complemento suficiente, son incapaces de destruir las células leucémicas. Cuando se les proporciona complemento (con el plasma perfundido), las células leucémicas son destruidas. Mi colega, William D. Hardy, Jr., ha demostrado que en los gatos también se produce esta destruc-



EFECTO ANTITUMORAL DEL BCG (bacilo Calmette-Guérin), una cepa atenuada del microorganismo de la tuberculosis. A los ratones del ensayo se les inyectó BCG y, una semana después (día cero en la gráfica), se sometieron a un trasplante tumoral; los controles recibieron únicamente el trasplante. En los animales infectados con BCG los tumores crecieron más lentamente o involucionaron totalmente *(curvas de color)*, mientras que en los controles crecieron de una forma progresiva hasta llegar a matar a todos los animales *(curvas negras)*. Parece que el bacilo calmette-Guérin funciona estimulando de modo general el sistema inmunológico del huésped.

LA NECROSIS HEMORRAGICA de un tumor trasplantado se produce tras la inyección de endotoxina, componente de la pared celular de las bacterias gramnegativas. El tumor se ennegrece al cabo de doce horas a causa de extensas hemorragias y de la muerte de las células *(centro)*. **La regresión ocurre varios días después** *(derecha)*. **El efecto de la endotoxina es indirecto y parece estar mediado por los macrófagos.**

ción de las células leucémicas, mediada por transfusión de plasma.

De la exposición de estos mecanismos de escape (y se han postulado varios otros), lo que debe quedar claro es que el crecimiento progresivo de un tumor no significa, en modo alguno, la ausencia de una respuesta inmunológica o de antígenos específicos respecto de ese tumor. Por descontado que el caso límite de mecanismo de escape sería la falta total de antígenos tumorales específicos. Se sabe de un proceso, llamado inmunoselección, en el cual las células más antigénicas de un tumor son destruidas por reacciones inmunológicas, persistiendo las células con menos antigenicidad que siguen multiplicándose, pero todavía no se ha hallado ningún tumor antigénico en el que se haya demostrado una pérdida permanente de antígenos tumorales específicos. Pero dadas las distintas formas en que los tumores pueden escapar a la destrucción inmunológica no es posible saber si algún tumor del ratón carece de antígenos tumorales. Actualmente tampoco se conoce una manera de comprobar el grado de antigenicidad de los antígenos tumorales específicos en el hombre, porque las técnicas de trasplante usadas para esa valoración en el ratón no pueden emplearse en el hombre. Mientras no se desarrollen las técnicas para obtener esta información, solamente se puede especular sobre si los modelos adecuados para el cáncer humano son los tumores del ratón con antígenos más potentes o aquellos otros con antígenos más débiles.

Inmunoterapia del cáncer

El descubrimiento de antígenos específicos de los tumores ha despertado naturalmente gran interés sobre la posibilidad de un enfoque inmunológico del tratamiento y prevención del cáncer. Resulta popular mencionar la inmunoterapia como la cuarta clase de tratamiento del cáncer (además de la cirugía, la radioterapia y la quimioterapia), pero sería un error inferir que existe, en este momento, ningún régimen de inmunoterapia comparable a los demás tratamientos, y que produzca efectos beneficiosos previsibles para el enfermo canceroso.

Los dos tratamientos anticancerosos de uso más general, la cirugía y la radioterapia, tienen el máximo efecto cuando el cáncer está localizado. Cuando las células cancerosas están diseminadas por todo el cuerpo, se necesita un tratamiento que pueda alcanzarlas y destruirlas donde sea que estén, y ésta es la principal ventaja teórica de la quimioterapia. A pesar de que ha habido un progreso muy significativo en el control de ciertas formas de cáncer por quimioterapia, todos los agentes químicos anticancerosos dañan, en mayor o menor grado, a las células normales, y esto suele poner límites a su eficacia final. El principal atractivo de la inmunoterapia radica en la extraordinaria especificidad de las reacciones inmunológicas. La destrucción de células cancerosas caracterizadas por antígenos específicos debería dejar intactas a las células normales que carecen de tales antígenos. Por esta razón se está trabajando muchísimo en laboratorios y clínicas de todo el mundo para aplicar principios inmunológicos al tratamiento de los procesos malignos.

Actualmente, se considera en general que los primeros intentos de inmunoterapia fueron experimentos de la naturaleza. Los médicos vienen observando, desde hace tiempo, casos en los que cánceres avanzados involucionan después de una infección bacteriana aguda. William B. Coley, un cirujano de Nueva York, nacido en el año 1862, dedicó su vida al desarrollo de una terapéutica del cáncer basada en esas observaciones. Coley empezó intentando infectar enfermos cancerosos con bacterias, pero, debido a que esto resultaba sorprendentemente difícil de conseguir, comenzó inyectando a los pacientes mezclas de bacterias muertas. Estas vacunas a base de mezclas de bacterias, que se conocieron con el nombre de toxinas de Co-

ley, reproducían algunas de las características de las infecciones bacterianas agudas, como fiebre y escalofríos, y, en un número significativo de casos, los tumores de sus pacientes involucionaron parcial o totalmente.

Helen Coley Nauts (hija de Coley), del Cancer Research Institute, ha reunido y analizado unas mil historias clínicas de pacientes tratados de esta forma, y no hay duda de que, en algunos, el tratamiento resultó muy eficaz. ¿Por qué cayó, pues, en desuso y, por algún tiempo, incluso en descrédito? Parte de la respuesta está en las grandes esperanzas que acompañaron a la aparición de la radioterapia hacia los años 1920 y 1930. Otra razón fue que no se podía ofrecer ningún razonamiento científico para explicar la actividad de las toxinas de Coley, y los médicos son reacios a adoptar una terapéutica cuyos fundamentos no entiendan y cuya eficacia no puedan predecir. A medida que progresó la inmunología, el avance de los conocimientos topó con el fenómeno de Coley y varios hallazgos empezaron a explicar el modo en que las toxinas de Coley ejercían su efecto antitumoral. Por un lado, se descubrió que los productos bacterianos ejercían una profunda influencia en las reacciones inmunológicas. Cuando los animales reciben un antígeno junto con bacilos tuberculosos muertos, fabrican mucho más anticuerpo que cuando reciben el antígeno solo. Con el descubrimiento de los antígenos específicos de los tumores en los animales, parecía razonable pensar que los tumores humanos fueran también antigénicos, y que los productos bacterianos de las toxinas de Coley reforzaran una respuesta inmune contra el cáncer que era insuficiente por sí sola.

Inmunopotenciadores

El último eslabón entre el trabajo de Coley y la era moderna se estableció cuando Baruj Benacerraf, que trabajaba a la sazón en la Escuela de Medicina de la Universidad de Nueva York, Donald A. Clarke y yo, encontramos que los ratones infectados con bacilo Calmette-Guérin (una cepa de bacilo tuberculoso atenuado que se emplea para vacunar a las personas contra esta enfermedad), mostraban una mayor resistencia a una serie de tumores. El bacilo Calmette-Guérin (BCG) produce en el ratón una infección transitoria y, en el curso de la reacción del animal a esta infección, su sistema inmunológico resulta estimulado al máximo. Los ratones infectados con BCG, además de ser más resistentes al cáncer, presentan una mayor capacidad para fabricar anticuerpo, rechazar injertos de tejido extraño normal y sobrevivir a algunas infecciones bacterianas y víricas que, normalmente, son letales. Se ha progresado considerablemente en la extracción e identificación de los componentes inmunológicamente activos del BCG, especialmente por parte de Edgar Lederer, del Instituto para la Química de las Sustancias Naturales, de Francia. Se han encontrado otros agentes cuyos efectos sobre la resistencia del animal receptor semejan los del BCG; algunos son de origen microbiano y otros son sintéticos.

Puesto que estos agentes, bajo ciertas condiciones, aumentan la reactividad inmunológica, han recibido el nombre de inmunopotenciadores.

Una de las principales metas de los inmunólogos tumorales consiste en determinar cómo funcionan los inmunopotenciadores; es decir, distinguir el papel respectivo de los componentes humorales y celulares del sistema inmunológico. Se ha puesto especial interés en el estudio del macrófago, como célula clave en la mediación del efecto del BCG y de otros inmunopotenciadores. El macrófago recibe el nombre de célula depredadora, por su capacidad para fagocitar (englobar) y digerir bacterias, células muertas y demás materiales extraños. El macrófago, desde que fue reconocido por Elie Metchnikoff a finales del siglo XIX, se ha venido considerando esencial para la salud, especialmente por su papel en la resistencia a la infección. El BCG y otros inmunopotenciadores tienen un notorio efecto sobre los macrófagos: al hacer que sean más fagocíticos, sintetizan más enzimas y se dividen. Peter Alexander, del Chester Beatty Research Institute, de Inglaterra, y Jack S. Remington, de la Palo Alto Medical Research Foundation, encontraron que estos macrófagos activados inhiben las células cancerosas en cultivo de tejidos, y algunas veces las matan, en condiciones en que las células normales no sufren daño.

Nuestro análisis de un fenómeno conocido de los investigadores del cáncer, desde hace más de 50 años, con el nombre de necrosis hemorrágica del tumor, ha contribuido a la comprensión del mecanismo por el cual los inmunopotenciadores pueden provocar la destrucción de las células tumorales. Algunos tumores del ratón y otros roedores sufren hemorragia aguda y destrucción después de la inyección de bacterias gramnegativas o de factores extraídos de las paredes celulares de estas bacterias. Esta reacción ha sido considerada durante mucho tiempo como el equivalente experimental de las observaciones de Coley en el hombre. El factor de la pared celular, causante de la necrosis hemorrágica del tumor y responsable también de muchos de los efectos tóxicos de estas bacterias, se llama endotoxina; químicamente se le ha identificado como un lipopolisacárido, un complejo de lípido (grasa) y moléculas de hidrato de carbono. A diferencia del BCG, que en los ratones es más eficaz si se administra antes de trasplantar un tumor, el efecto de la endotoxina se hace más intenso sobre tumores establecidos.

La acción de la endotoxina no es directa, como lo demuestra el hecho de que *in vitro* no sea letal para células tumorales en cultivo de tejidos. Durante muchos años se creyó que la necrosis hemorrágica de los tumores era el resultado de una acción de la endotoxina sobre el sistema circulatorio, que consistía en inducir una disminución de la irrigación sanguínea del tumor y ocasionar así la muerte de las células por falta de oxígeno. El trabajo de nuestro grupo en el Sloan-Kettering Institute ofrece otra explicación para este efecto de la endotoxina. En la sangre de los animales a los que se ha inyectado endotoxina hemos detectado un factor capaz de matar directamente las células tumorales. Nuestros resultados actuales nos llevan a pensar que la endotoxina hace que los macrófagos liberen este factor, el que a su vez es responsable de la destrucción de las células tumorales.

En este momento se están descubriendo inmunopotenciadores que actúan selectivamente sobre ciertos componentes del sistema inmunológico, en contraste con el BCG y otras sustancias microbianas que actúan de modo más amplio. Posiblemente, a medida que aprendamos a medir y a entender los defectos inmunológicos específicos que se asocian al cáncer o a otras enfermedades, estos inmunopotenciadores más específicos se convertirán en valiosos agentes terapéuticos.

La inmunopotenciación, tanto si es general como selectiva, no constituye más que una de las muchas formas de aplicar principios inmunológicos al tratamiento del cáncer. La idea de una vacuna específica para el cáncer es tan vieja como el mismo pensamiento inmunológico. Los médicos han intentado innumerables veces variar el curso del cáncer inyectando un extracto de células cancerosas del propio paciente, o de otro enfermo canceroso. Teniendo en cuenta que la mayoría de estos sujetos estaban en una fase avanzada de su enfermedad, cualquier resultado beneficioso hubiese sido sorprendente y, de hecho, no lo hubo. Sin embargo, con

los nuevos conceptos y hallazgos de la inmunología del cáncer, la inmunización específica con vacunas tumorales es una posibilidad atractiva que tiene una base científica racional. Mas, para que este enfoque sea eficaz, necesitamos conocer mejor la fuente y forma más adecuadas de antígeno tumoral y encontrar la manera más eficaz de combinarlo con los inmunopotenciadores. Como de costumbre, el trabajo en el ratón llevará seguramente la delantera en cuanto a proporcionar la información necesaria.

Futuros caminos

El mayor triunfo de la investigación inmunológica ha sido su éxito en la prevención de enfermedades infecciosas que antes eran devastadoras para la humanidad. El futuro dirá cuál pueda ser su contribución a la comprensión y control del cáncer. La disciplina de la inmunología del cáncer es muy nueva, y no sería realista esperar en este momento grandes aplicaciones prácticas de sus principios. Lo que podemos esperar en estos próximos años es un mayor conocimiento de los antígenos tumorales específicos en los animales experimentales y un estudio detallado de la respuesta inmune que inducen. El estudio de la inmunología se ha visto profundamente afectado por el descubrimiento de que los genes ejercen un rígido control sobre la naturaleza y la intensidad de las reacciones inmunológicas; ahora se necesita definir el control genético de la respuesta inmunológica al cáncer y a los virus cancerígenos. La mejor manera de evitar que los tumores escapen a la destrucción inmunológica quizá consista en averiguar la forma de modificar o vencer las restricciones genéticas de la respuesta inmune. En la investigación sobre el cáncer humano, la tarea principal reside en aplicar los hallazgos derivados del estudio de los tumores animales. Entre otras cosas, esto supone conseguir estimas cuidadosamente controladas del valor de la inmunopotenciación en el tratamiento del cáncer humano, así como realizar rigurosos estudios de laboratorio para definir los antígenos del cáncer humano con la misma precisión con que podemos hacerlo en los animales de experimentación.

Mirando al futuro inmediato, no parece que se justifiquen ni un optimismo irreflexivo ni nuevas olas de duda y pesimismo. Se han registrado avances importantes y puede esperarse que la investigación inmunológica siga aportando notables contribuciones a nuestra comprensión de las enfermedades malignas.

La desaparición de las manchas solares

Documentos antiguos indican que entre 1645 y 1715 no hubo prácticamente manchas en el sol. Parece probable que la actividad de éste varíe de forma considerable y que el periodo actual sea un periodo de actividad poco corriente

John A. Eddy

En 1893 E. Walter Maunder, superintendente de la sección solar del Royal Greenwich Observatory de Londres, buscaba entre libros y periódicos antiguos y casi no podía creer lo que estaba descubriendo. Al parecer, durante muchos años los astrónomos no se habían dado cuenta de una realidad muy importante: el sol no era la estrella regular y predecible que todos habían supuesto. Si Maunder podía aceptar lo que estaba leyendo, el sol había cambiado en gran medida en épocas bastante recientes. Concretamente, las referencias antiguas revelaban que durante un periodo de 70 años, que terminaba alrededor del 1715, las manchas y toda otra actividad solar sin excepción habían desaparecido del sol. De ocurrir lo cual, Maunder sabía que las consecuencias podían ser muy importantes, no sólo para la astronomía sino quizás también para el clima y, por consiguiente, para las futuras condiciones de vida sobre la tierra.

Las manchas solares constituyen el fenómeno solar mejor conocido; su ciclo de unos 11 años, en el que aparecen y desaparecen, figura entre los hechos astronómicos mejor documentados. Aun cuando, durante siglos, los astrónomos de Oriente habían registrado grandes manchas solares observadas a simple vista, en el mundo occidental se pasó en general por alto las manchas oscuras del sol, hasta que en 1611 Galileo y otros las observaron con el telescopio. Desde entonces hasta el presente se han visto sometidas a una continua vigilancia telescópica. En 1843 Heinrich Schwabe, astrónomo aficionado alemán, advirtió con los datos recogidos por él mismo que, cuando representaba gráficamente el número medio de manchas solares observadas por año, aparecía una nítida variación cíclica con un periodo de unos 10 años. Su descubrimiento sorprendió a los astrónomos profesionales que habían sostenido la no existencia de ningún ciclo ni en la aparición de manchas ni en otra actividad solar. Sin embargo, poco después de la declaración de Schwabe, otros observadores confirmaban la realidad del ciclo, ajustándolo a 11,2 años. A partir de referencias de antiguos observatorios, el astrónomo suizo Rudolf Wolf estableció también que el ciclo había estado transcurriendo por lo menos desde 1700, fecha que le parecía que era el límite de fiabilidad de las referencias de que disponía. En 1893 era bien sabido que el ciclo de las manchas solares se encontraba asociado a otros signos de actividad solar y a fenómenos terrestres periódicos tales como las auroras. La curva del número anual de manchas solares desde 1700 parecía parte de una onda que se prolongara de forma invariable en el tiempo tanto hacia adelante como hacia atrás. Pocos dudaban de que las manchas y el ciclo undecenal fuesen características perdurables del sol. Entonces como ahora, estas características se aceptaban como pruebas de que la actividad solar era regular, lo cual a su vez significaba que el sol era constante y pronosticable.

En 1893, el sol estaba en un máximo del ciclo de las manchas solares, pudiéndose ver cientos de ellas, como Maunder sabía muy bien. Incluso en los años en que el ciclo está en un mínimo, suelen darse, cuando menos, unas cuantas manchas: es raro que pase un mes entero sin que aparezca ninguna mancha solar. No obstante, en los viejos documentos del siglo XVII, precisamente antes del comienzo de la familiar curva de manchas solares, Maunder había encontrado relatos de fuentes originales acerca del paso de año tras año sin la aparición de manchas. A lo largo de 32 años no se había visto ni una mancha en el hemisferio norte solar. Durante 65 años sólo se registró un pequeño grupo de ellas alguna vez. Habían transcurrido periodos hasta de 10 años de duración sin que nadie encontrara una mancha en ningún lugar del sol. Maunder observó que el número total de manchas reseñadas entre 1645 y 1715 era menor que el que se ve actualmente en un solo año.

Maunder publicó en 1894 un trabajo titulado *A Prolonged Sunspot Minimum* (Un prolongado mínimo de manchas solares): en él aportaba detalles a propósito de ese raro periodo en la historial del sol y llamaba la atención sobre sus consecuencias. Si la aparente escasez de manchas era real, ello debería sacudir los fundamentos de la astronomía solar. Maunder indicó que tal periodo extraordinario podría servir como una eficaz comprobación de las relaciones entre la tierra y el sol. En efecto: si las elevaciones y depresiones normales del ciclo undecenal de manchas solares podían detectarse en las variaciones en el campo magnético de la tierra o quizás en el clima, entonces un cambio prolongado en el comportamiento del sol debía acompañarse de efectos importantes en la tierra.

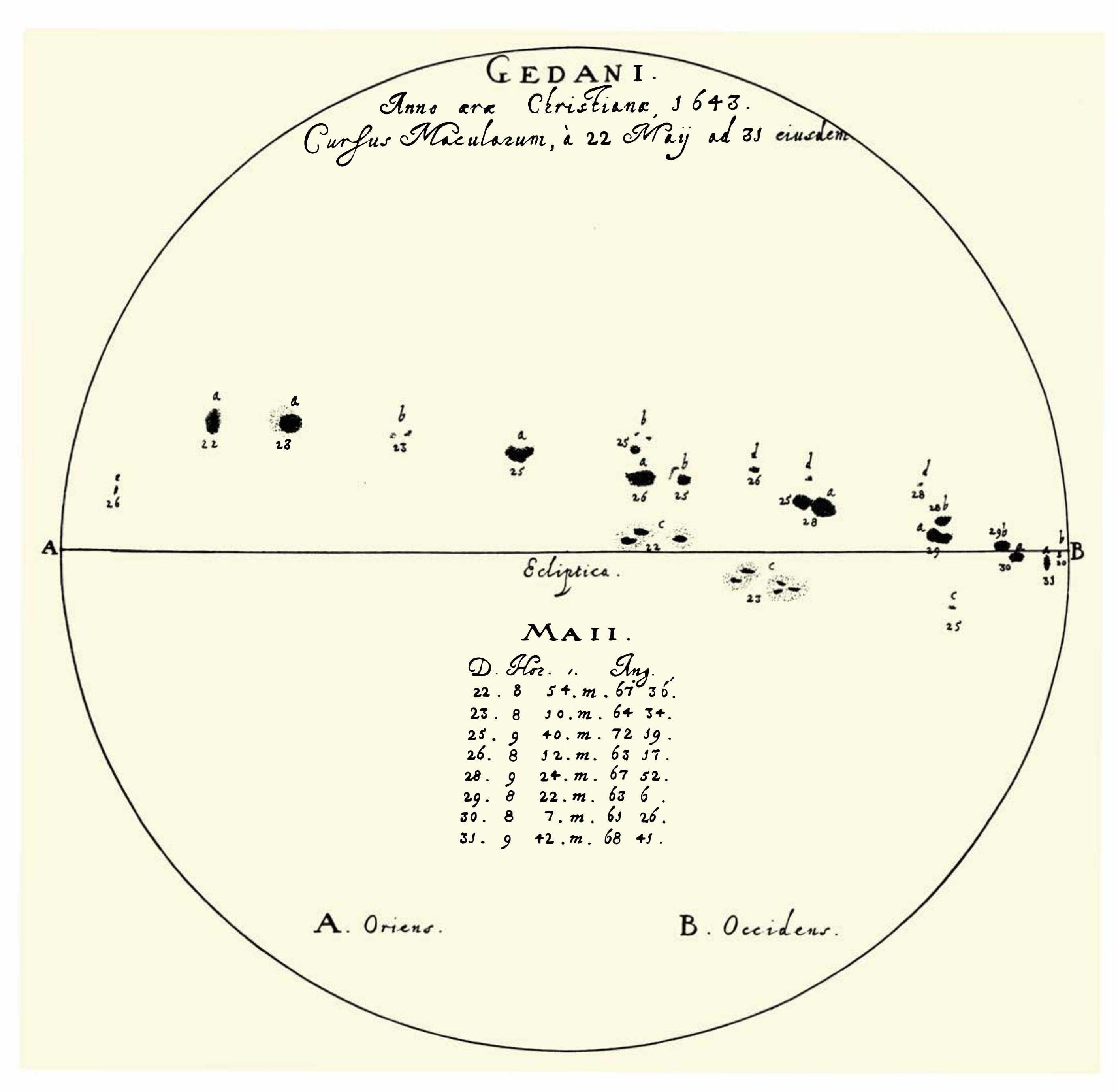
No se sabe si alguien prestó oídos a Maunder. Tampoco se había concedido demasiada atención a un trabajo que había escrito sobre el mismo tema en 1890, o a otro publicado un año antes por el astrónomo alemán Gustav Spörer; éste fue quien hizo caer a Maunder en la existencia de un periodo falto de manchas solares.

Pero Maunder no renunció. En 1922 volvió de nuevo sobre el tema con otra publicación titulada también *A Prolonged Sunspot Minimum,* recalcando asimismo la importancia del periodo de 70 años para la astronomía solar y la física terrestre. Seis años más tarde moría Maunder; el ciclo de las manchas solares siguió sin cesar como si se mo-

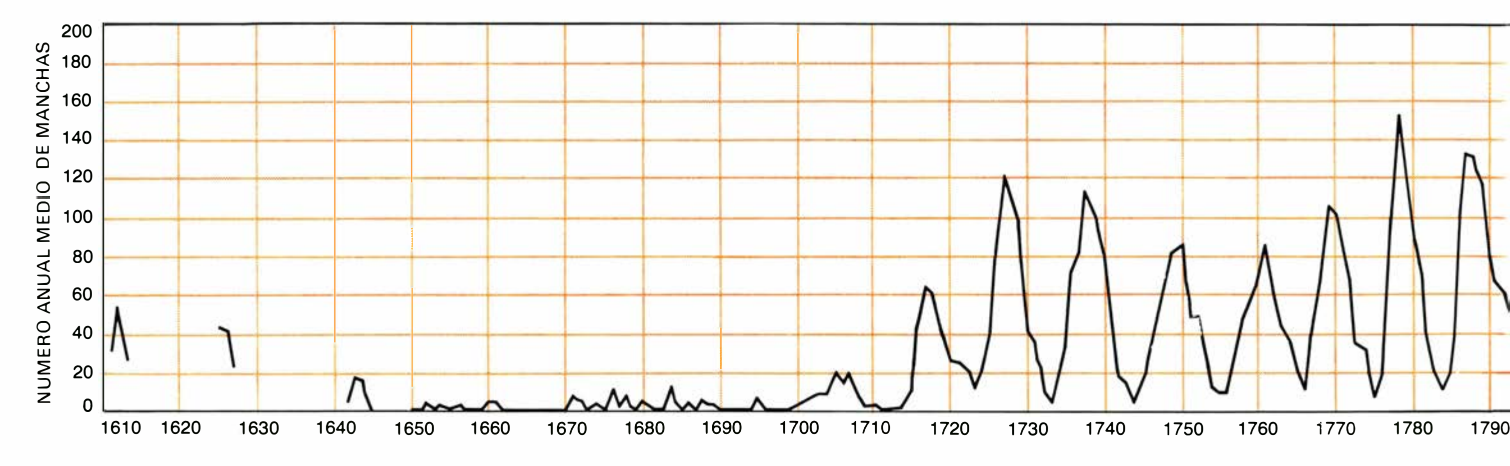
fara de él. Sus publicaciones cayeron en el olvido o se atribuyeron a su talante entusiasta, que tenía demasiada fe en registros antiguos y fragmentarios.

Varios años después me pareció que había llegado el momento de poner en claro el caso de la desaparición de las manchas solares, que, durante un tiempo excesivo, había estado colgado cual un esqueleto en el armario de la física solar. Me desazonaban las referencias ocasionales a él en relación con un cambio coincidente en el clima del mundo. Como astrónomo solar, me parecía que nunca podía haber sucedido, y mi interés por la historia hizo que la perspectiva de verificar las aserciones de Maunder me resultase una tarea atractiva.

El trabajo tenía la intriga de las novelas policiacas. Se había cometido un crimen, grave para la astronomía y quizás para la misma tierra, según las informaciones recogidas hacía de ello mucho tiempo. ¿Sucedió realmente? Las pistas originales de Maunder sobre el caso tenían ahora más de 250 años de antigüedad, y, sin embargo, estaban todavía intactas en las bibliotecas que conservaban diarios de los siglos XVII y XVIII. Resultaba aún más estimulante el que en el medio siglo transcurrido desde la muerte de Maunder hubiesen aparecido nuevas pruebas en el adelanto de la física solar. Entre estas pruebas había catálogos de observaciones históricas de auroras, compilaciones de manchas solares observadas a simple vista en Oriente, y una comprensión más completa de cómo podría aparecer un sol



DIBUJO DE MANCHAS SOLARES del siglo XVII, publicado por Johannes Hevelius de Danzig en su libro *Selenographia* (1647). Muestra la evolución de varios grupos de manchas desplazándose sobre el disco del sol, que está girando, desde el 22 de mayo de 1643 hasta el 31 de mayo del mismo año. A partir de una serie de dibujos parecidos, el autor y sus colaboradores han deducido la velocidad de rotación del sol a mediados del siglo XVII, precisamente al comienzo del periodo de 70 años en que casi no hubo manchas. La forma de las manchas en el dibujo y los detalles de la oscura sombra de las mismas circundada por la penumbra, más brillante, demuestran que la calidad de los telescopios del siglo XVII era tan buena que los astrónomos de entonces podían observar las manchas con casi igual detalle que los astrónomos actuales.



EL CICLO DE LAS MANCHAS SOLARES no es regular ni en frecuencia ni en amplitud, como puede verse en esta curva del número anual medio de manchas, una medida del número de manchas visibles en el disco del sol en cualquier época desde 1610 hasta 1976. El intervalo entre dos máximos consecutivos del ciclo de las manchas no es de 11 años de manera regular; los hay más cortos, de 8 años, y otros más largos, de hasta 17 años. Además, algunos máximos tienen mayor importancia que otros, como puede verse comparando el de 1959 con los

absolutamente inactivo durante un eclipse solar total. Lo mejor de todo era que tenía a mi disposición una poderosa herramienta en el moderno análisis de los anillos de crecimiento anual de los árboles. Lo que a mí me importaba no era la anchura de los anillos, que evalúa sólo el clima local, sino su contenido químico, que comporta un registro indirecto de los cambios acaecidos en el sol. Maunder no había conocido ninguna de estas pistas.

Ninguna de ellas era por sí misma concluyente. Sin embargo, todas ellas en conjunto, más las referencias históricas originales, podrían a buen seguro resolver el inquietante caso.

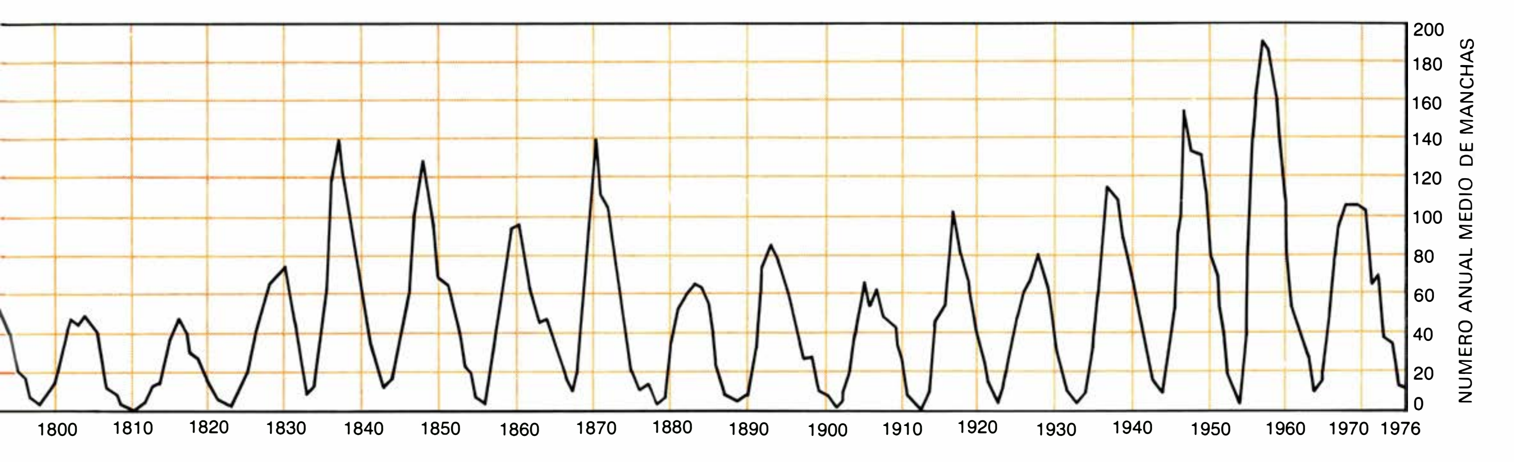
Estudiando los informes de los astrónomos de ese periodo encontré, con sorpresa por mi parte, que eran exactamente como Maunder los había representado; en su honor, empecé a llamar al periodo falto de manchas el "mínimo de Maunder". Cuando en 1671 se registró una nueva mancha solar, en medio del mínimo de Maunder, el editor de *Philosophical Transactions of the Royal Society of London* dijo: "En París el admirable Sr. Cassini ha detectado otra vez, recientemente, manchas en el sol, de las cuales ni una había sido vista en muchos años, que nosotros sepamos".

El editor continuaba describiendo la última mancha vista 11 años antes para los lectores que hubieran olvidado a qué se parecía una mancha solar. El mismo Cassini (G. D. Cassini, fundador y primer director del Observatorio de París) escribió a propósito de dicho fenómeno: "Se van a cumplir los veinte años desde que los astrónomos vieran la última mancha importante en el sol; antes, desde la invención del telescopio, se habían detectado de vez en cuando". ¿Podían estas palabras haber sido escritas en una época de comportamiento normal del sol?

John Flamsteed, primer astrónomo real de Inglaterra y primer director del Royal Greenwich Observatory, dejó constancia de cómo exploró el sol durante siete años antes de que pudiera encontrar ni una sola mancha. En los libros escritos bastante después del final del periodo falto de manchas solares se hacían constantes referencias al mismo. De una manera significativa no fue sino hasta el descubrimiento por Schwabe del ciclo de 11 años cuando empezaron a desaparecer las referencias a la larga ausencia de manchas en el siglo XVII. Una ausencia continuada de manchas podría explicar por qué los astrónomos se sorprendieron en principio por el anuncio del ciclo de Schwabe. Un ciclo de manchas solares suprimido o interrumpido podría explicar por qué transcurrieron más de 230 años entre la primera observación telescópica de manchas solares y el tardío descubrimiento de que éstas aparecen y desaparecen a lo largo de un ciclo pronunciado.

Pero, ¿podemos fiarnos de los observadores antiguos? ¿Eran buenos sus telescopios? ¿Con qué rigor observaron las manchas solares? El siglo XVII no fue anteayer. Era la época de Luis XIV y la gente llevaba vestidos primorosos y escribía con rebuscamiento. Sin embargo, en el mismo periodo Cassini descubrió la mayor división oscura de los anillos de Saturno y se estableció que éste tenía por lo menos cinco satélites. El mínimo de Maunder empezó 35 años después de que Galileo construyese su primer telescopio. En los años que mediaron, florecieron tanto la óptica como la astronomía. En el siglo XVII se desarrollaron los telescopios de suspensión que tenían una distancia focal de hasta 60 metros. Fue la época del primer telescopio reflector de Newton y de otras muchas innovaciones. Los astrónomos observaron y contaron manchas en el sol casi de la misma manera en que lo hacemos hoy en día, y sus instrumentos apenas diferían de los empleados para tal fin en los dos siglos siguientes. Sus dibujos de las manchas solares, conservados en publicaciones periódicas y libros, no desmerecen de los dibujos realizados por los observadores de 1977. Estoy convencido de que los astrónomos del siglo de Luis XIV poseían los instrumentos y la habilidad necesarios para ver todas las manchas solares, excepto las más diminutas, si hubiesen aparecido manchas en el sol. Creo que todos estos observadores eran tan capaces como nosotros, que muy probablemente tenían los mismos estímulos y el mismo interés profesional y que seguramente disponían de más tiempo para emplear en el telescopio.

¿Mantuvieron al sol en continua vigilancia? ¿Tomó Maunder la ausencia de pruebas por prueba de la ausencia? Dos hechos me sugieren que el problema no es la ausencia de pruebas. La ausencia de manchas solares se observó repetidamente durante el periodo; si se acepta que los motivos de los observadores del siglo XVII fueron muy parecidos a los nuestros, sospecho que se debió mantener una vigilancia especialmente estrecha para descubrir nuevas manchas, para contradecir lo que incluso entonces se advertía como algo extraño. Además, hay artículos en las publicaciones de aquella época que indican que el descu-



de principios del siglo XIX. Al periodo entre 1645 y 1715 el autor lo denomina el mínimo de manchas solares de Maunder, en honor del físico solar británico E. Walter Maunder, que fue el primero en señalar la significación del período para las condiciones en la tierra. El registro de manchas observadas antes de 1650 es fragmentario. El primer máximo aislado, alrededor del 1612 *(izquierda)* se extrajo de las observaciones de Galileo; el segundo, de las observaciones de Christoph Scheiner, descritas en su libro *Rosa ursina;* el tercero, de las observaciones de Hevelius.

brimiento de una nueva mancha proporcionaba la ocasión de publicar un informe. Actualmente, hasta en el mínimo de manchas solares se ven tantas manchas que si se escribiera un informe por cada una de ellas, ninguna publicación podría editarlas todas.

¿Pudo el tiempo meteorológico haber interferido las observaciones? ¿Pudo tener Europa un número excepcionalmente grande de días nublados durante 70 años que hubiese mantenido los astrónomos alejados de sus telescopios? Es verdad que se registró un frío intenso en Europa, pero no fue un tiempo de cielo totalmente cubierto. Si lo hubiese sido lo sabríamos por las lamentaciones, en las publicaciones de los astrónomos, que nunca se han distinguido por ser ni muy pacientes ni muy taciturnos. Además, en el siglo XVII la astronomía nocturna se mostró vigorosa y activa: se descubrieron cometas regularmente, y los adelantos de la época en el conocimiento de los planetas necesitaron no sólo un cielo despejado sino también una atmósfera estable.

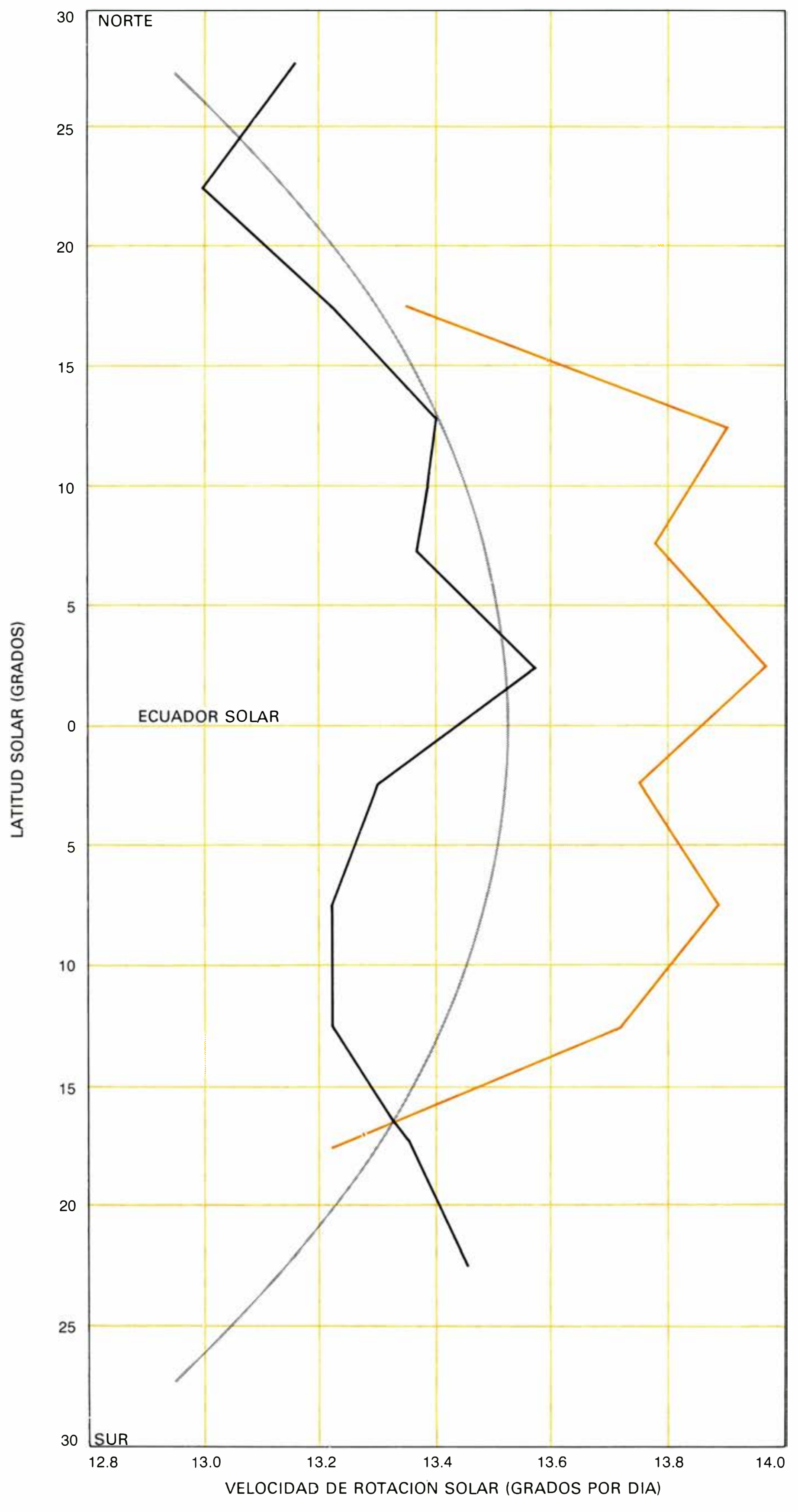
Las descripciones históricas de la aurora boreal permiten aun menos dudas respecto de la realidad del mínimo de Maunder. Las manifestaciones aurorales están ligadas al nivel de actividad solar. Por debajo del Círculo Artico, el número de noches en que se ve una aurora está en correlación con el número de manchas que hay en el sol. En general la frecuencia con que se ven las auroras depende también de la distancia del observador a los polos magnéticos de la tierra. Las auroras son más frecuentes en latitudes más altas y escasean cerca del Ecuador porque, a latitudes inferiores, la geometría de las líneas de fuerza del campo magnético de la tierra ampara la atmósfera contra las partículas que provienen del sol y que causan las auroras boreales.

En condiciones ordinarias de actividad solar normal, en las populosas áreas de Europa se observan por lo menos 500 y quizás hasta 1000 auroras en un periodo de 70 años. Sin embargo, entre 1645 y 1715 apenas si se registraron auroras en Europa. Incluso en Escandinavia, donde actualmente se ven auroras casi todas las noches, se registraron tan pocas que se las consideró un portento. Durante el mínimo de Maunder hubo un intervalo de 37 años en que no se observó ni una sola aurora en ningún lugar de la tierra. Cuando finalmente se vio una en Inglaterra, en marzo de 1717, al final del mínimo de Maunder, el astrónomo Edmund Halley, que era entonces astrónomo real, se animó a escribir un artículo explicando el fenómeno. Confesó que nunca había visto ninguna aurora, aunque tenía entonces 60 años y siempre había estado al acecho. Aunque Halley no lo supo, su vida había transcurrido durante la mayor parte del mínimo de Maunder.

Los 70 años del mínimo de manchas solares es una característica particularmente bien definida si se hace la gráfica del número de referencias a auroras históricas por año. Con solo mirar tales referencias, Maunder podía igualmente haber escrito sus artículos. En el recuento de las auroras históricas, sin embargo, hay también otro aspecto que exige explicación. Tiempos muy atrás, se informó de un número pequeño de auroras en comparación con el recuento que hoy se hace. ¿Por qué tales informes fueron tan raros hasta llegar a 1645? Los documentos muestran que el número de auroras descritas empieza a elevarse por el año 1550, se interrumpe por el mínimo de Maunder y entonces salta con un factor de por lo menos 20 después de 1716. ¿Qué proporción de este notorio aumento en la cifra de auroras descritas después de 1550 es un efecto social, es decir, un reflejo del interés por la astronomía en el Renacimiento o, en una época más reciente, la repercusión de la publicación de Halley?

Sospecho que gran parte de este aumento del número de auroras descritas desde el Medievo es verdaderamente un efecto social. Otra prueba, sin embargo, me hace pensar que por lo menos una fracción de este incremento es un efecto físico, a saber, el efecto de un sol que cambia. Hay pruebas de que hubo otros largos periodos parecidos al del mínimo de Maunder en tiempos más remotos. Se muestran claramente en primitivas descripciones de auroras y en la frecuencia de descripciones de manchas solares observadas a simple vista. He llegado a creer que, mirado a largo plazo, la frecuencia actual de manchas y auroras es probablemente insólita, y que, desde el siglo XVII, la actividad del sol ha aumentado constantemente hasta alcanzar un nivel altísimo, un nivel quizás sin igual en este milenio.

Las descripciones de las manchas solares observadas sin la ayuda de un telescopio sirven como confrontación de las descripciones de las auroras y de la realidad del mínimo de Maunder. Existen descripciones de manchas solares que datan por lo menos del siglo quinto antes de Cristo y a partir de entonces se las registró con bastante regularidad, especialmente en Oriente. Las grandes manchas, o los grupos extensos de ellas, pue-



LA ROTACION DEL ECUADOR SOLAR se aceleró justamente antes del comienzo del mínimo de Maunder. El sol no gira como un cuerpo rígido; las regiones ecuatoriales giran más rápidamente que las regiones de latitudes más altas. La velocidad moderna de rotación del sol a distintas latitudes se ha dibujado en grados por día *(gris)*. Las observaciones de Scheiner muestran que la velocidad de rotación en 1620 era casi la misma que la actual *(negro)*. Las observaciones de Hevelius en 1640 *(color)* muestran que la velocidad de rotación ecuatorial del sol, comparada con la velocidad de rotación polar, había aumentado por un factor de tres. Sigue desconociéndose todavía si este incremento fue la causa o, por el contrario, un efecto del mínimo de Maunder.

den observarse fácilmente a simple vista mientras el sol está saliendo o se está poniendo, o cuando está muy oscurecido y coloreado por la niebla. En 1933 el astrónomo japonés Siguru Kanda compiló una lista de observaciones conocidas de manchas solares efectuadas a simple vista desde el Japón, China y Corea. Estableció que en la Era Cristiana se habían visto un promedio de entre 5 y 10 de tales manchas por siglo, incluyendo unos pocos periodos en que se vieron con más frecuencia y varias lagunas en que no se vieron manchas. Uno de tales vacíos se extendía desde 1584 a 1770 d. C., que comprendía el mínimo de Maunder.

Estas pruebas son fragmentarias y hay efectos sociales que podrían explicar dichas lagunas, en especial si se tiene en cuenta que el número normal de observaciones es tan pequeño. Yo estaría dispuesto a descartar las lagunas en las observaciones de manchas solares en Oriente, como coincidencias, si no fuera por el hecho de que las descripciones de las manchas solares observadas a simple vista se corresponden muy bien con la frecuencia de las auroras registradas en Europa durante un periodo de más de 2000 años.

Además de las informaciones sobre auroras y manchas solares, las descripciones del aspecto de la corona solar durante un eclipse total del sol proporcionan otro tipo de pruebas para verificar los niveles de actividad solar en el pasado. La forma de la corona, la tenue atmósfera más exterior del sol, es muy distinta cuando hay muchas manchas de la que tiene cuando hay pocas manchas. Las manchas solares son focos de intensos campos magnéticos en la superficie del sol y estos campos magnéticos dan forma a las ráfagas coronales de color blanquecino que se ven alrededor del sol durante un eclipse total. Cuando el número de manchas es grande, e intenso consiguientemente el campo magnético solar, la corona presenta muchas ráfagas que se prolongan a partir del astro como los pétalos de una dalia. Cuando hay menos manchas, el número de ráfagas coronales disminuye como si alguien hubiera arrancado los pétalos. En el mínimo de las manchas solares la corona que se ve durante un eclipse total es opaca y se limita a unas pocas ráfagas en las proximidades del ecuador.

¿Qué le sucedería a la corona si no hubiera manchas solares y no hubiese aparecido ninguna durante años o décadas? Un observador seguiría viendo la luz zodiacal o "falsa corona" alrededor del disco negro de la luna, un estrecho anillo de pálida luz rojiza, sin nin-

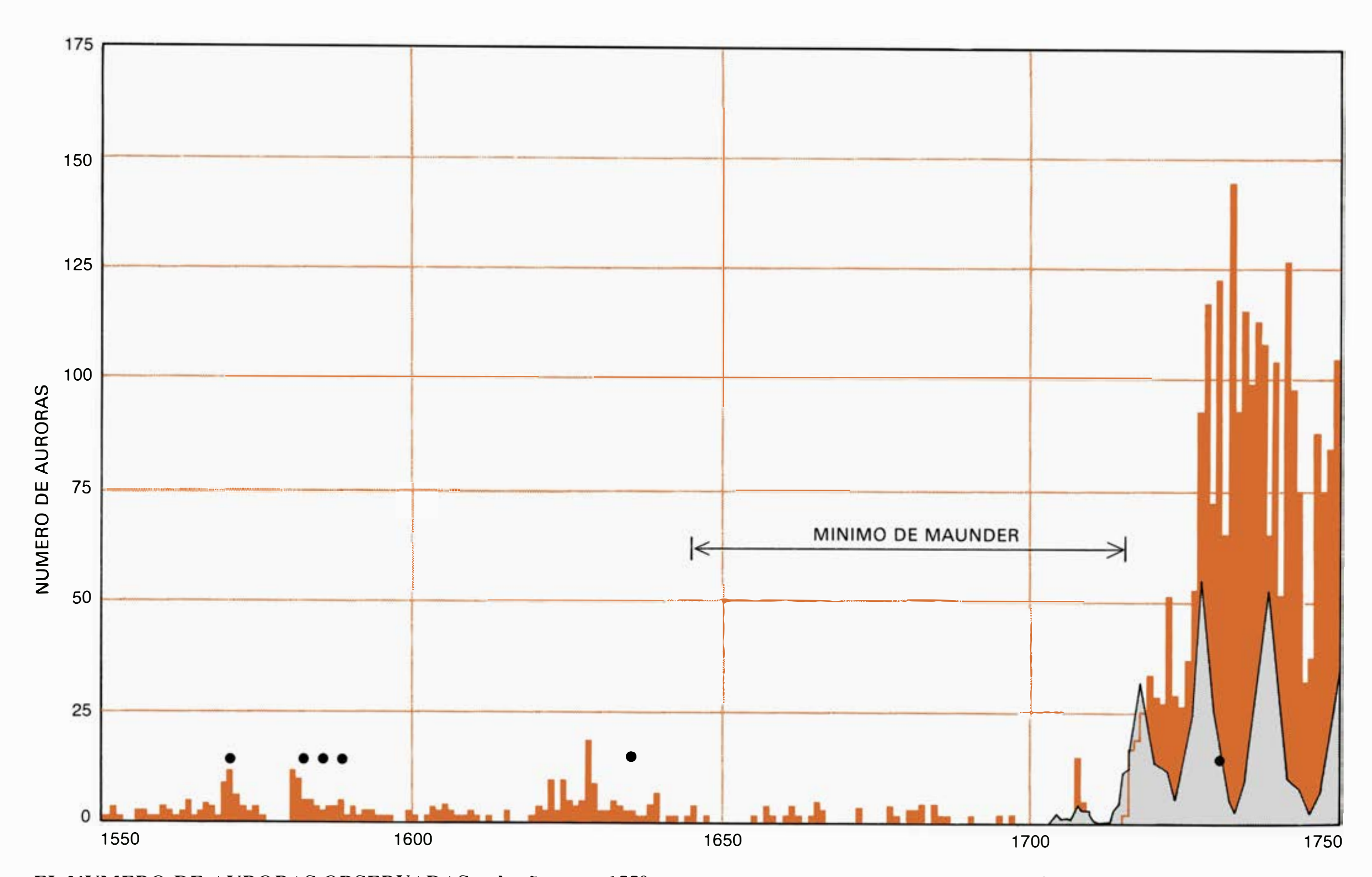
guna ráfaga. A diferencia de la verdadera corona, la luz zodiaca! no es una parte de la atmósfera solar; es simplemente luz solar difundida por el polvo interplanetario. Durante un eclipse total de sol, la luz zodiacal, aunque presente, está por lo común velada por el mayor brillo de la corona.

Entre 1645 y 1715 hubo 63 oportunidades de ver el sol en eclipse total. He procurado seguir la huella de las observaciones de todas ellas para hallar cómo describían la corona en la época del mínimo de Maunder. La mayoría de los eclipses se pudieron observar sólo desde regiones inaccesibles; en los siglos XVII y XVIII, no se acostumbraba, como hoy en día, enviar expediciones a seguir la sombra de la luna hasta los confines de la tierra. Algunos se contemplaron desde puntos cercanos a los observatorios europeos y éstos fueron particularmente bien observados. Ni un solo eclipse europeo fue desperdiciado; algunos más se registraron por parte de estudiosos que se encontraban por casualidad en el Nuevo Mundo o en Asia.

La mayoría de los astrónomos profesionales que observaron un eclipse no levantaron la vista al cielo para mirar el sol, sino que emplearon el precioso tiempo mirando la imagen del eclipse proyectada, a través de un telescopio, sobre una pantalla. Sobre la imagen calibraron los detalles del oscurecimiento del sol por la luna, que era entonces tema de urgente interés. Sin embargo, los que levantaron los ojos al cielo cuando el eclipse era total están de acuerdo en sus descripciones: en torno a la luna había un pálido anillo de luz tenue, de anchura uniforme, rojizo y estrecho. Ningún observador describió la blanquecina estructura coronal que es tan familiar y tan sorprendente hoy en día. Basándome en tales relatos, he llegado al convencimiento de que en la época del mínimo de Maunder la corona provista de ráfagas estuvo enteramente ausente o era tan tenue que todo lo que se podía ver era la luz zodiacal.

De acuerdo con ello, los informes sobre eclipses parecerían constituir una prueba clara de que el mínimo de Maunder existió realmente. Al igual que con las descripciones de las auroras, debemos también considerar que la corona se había visto antes de esta época. ¡No he podido encontrar ninguna descripción de una estructura coronal, obtenida durante algún eclipse, en ninguna época anterior al mínimo de Maunder! Los eclipses totales han fascinado la atención del hombre y le han infundido temor durante milenios. ¿Cómo, pues, en miles de descripciones de cientos de eclipses totales de sol, pudo haber escapado a su atención el espectáculo más bello y conmovedor de todos? Quizás vemos sólo lo que esperamos, por muy sorprendente que sea la escena: así, no se supo que la corona fuese parte del sol hasta finales del siglo XIX. Quizás los observadores quedaban momentáneamente deslumbrados por los últimos rayos del sol antes que el eclipse alcanzara la totalidad y por eso no veían la etérea corona. Quizás antes de la Ilustración, en el siglo XVIII, nadie se había sentido incitado a describir la corona.

Sospecho que a cualquiera que haya contemplado a simple vista el espléndido desarrollo de la corona, estas excusas le parecerán, como a mí, banales. Resulta verosímil que la corona haya estado realmente ausente del sol durante un periodo mayor que el mínimo de Maunder, posiblemente a lo largo de varios siglos. Su ausencia coincidiría entonces con un prolongado mínimo de la actividad solar que habría empezado antes, al que llamo mínimo de Spörer, que se evidencia en el



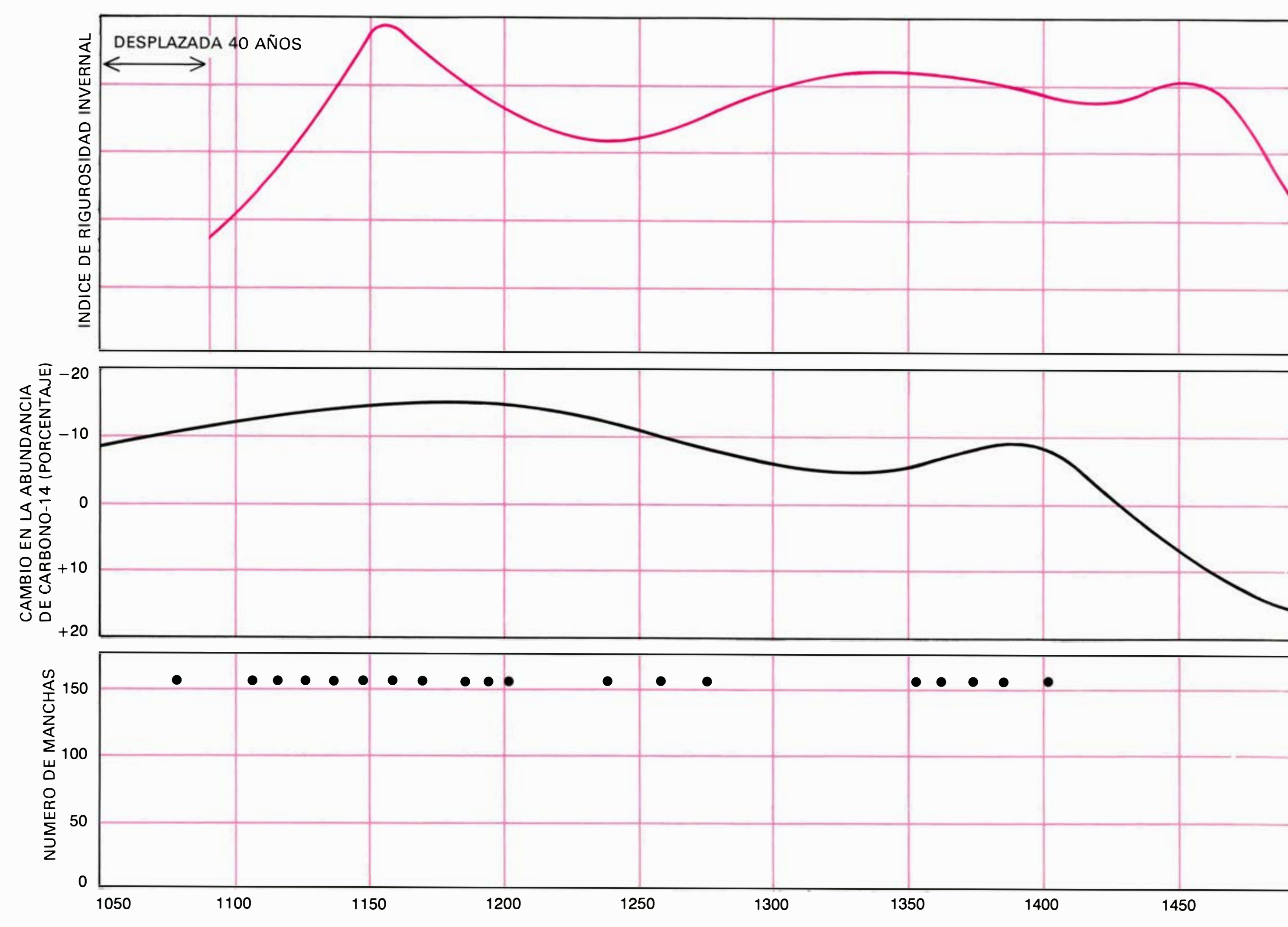
EL NUMERO DE AURORAS OBSERVADAS cada año, entre 1550 y 1750 *(espículas de color)*, muestra que las apariciones de las auroras boreales, que se encuentran relacionadas con el nivel de actividad solar, casi desaparecieron durante el mínimo de Maunder. A principios del siglo XVIII, cuando parece que empezó el ciclo moderno de manchas solares *(gris)*, el número de auroras se incrementó en 20 veces. Las épocas en que se vieron mayor número de auroras corresponden con las épocas enque se describieron en Oriente manchas observadas a simple vista *(puntos)*.

cómputo de auroras, en la ausencia de observaciones de manchas solares a simple vista y en el registro que proporcionan los anillos de los árboles. Tal ausencia implica que la corona de hoy día podría ser una característica transitoria y moderna del sol.

Otra información independiente con respecto al mínimo de Maunder proviene del trabajo que he llevado a cabo con Dorothy E. Trotter y Peter A. Gilman sobre la reconstrucción del modelo de la rotación solar en el siglo XVII. Observaciones modernas demuestran que el sol no gira como un cuerpo rígido; se retrasa en las latitudes más altas: el ecuador solar da una vuelta en 27 días y sus regiones polares la dan en 31 días. La moderna teoría solar sostiene que la rotación diferencial observada en el sol produce el ciclo de las manchas en virtud de la acción de una a modo de dinamo solar (como si el sol fuera una dinamo), asiento de un campo magnético solar profundo (el campo magnético forma parte de la dinamo y lo lleva en su interior, en las profundidades del sol) que interacciona con la rotación diferencial de la superficie para generar corrientes eléctricas, originando los campos magnéticos de las manchas solares. Por consiguiente, si hubiera una marcada anomalía en el número de manchas, se podría esperar una correspondiente anomalía en el campo magnético solar o en la rotación de su superficie.

La rotación del sol puede medirse con exactitud a partir de series de gráficos en los cuales se muestre la posición diaria de las manchas. Estas sirven de indicadores de la rotación superficial al igual que las hojas que flotan en una corriente trazan el fluir del agua. Con Trotter y Gilman hemos reconstruido los modelos de rotación solar en el siglo XVII examinando dibujos originales del sol y de las manchas solares de dos libros antiguos: *Rosa ursina*, de Christoph Scheiner, publicado en 1630, antes de que empezara el mínimo de Maunder y *Selenografía* de Johannes Hevelius, el cual apareció casualmente en 1647, al comienzo del mínimo.

En cada uno de estos libros hay dibujos del sol durante un periodo casi continuo de cerca de dos años. La rotación solar que deducimos de los dibujos de Scheiner en *Rosa ursina* para los años 1625 y 1626 se parecen mucho al modelo actual: la rotación del sol varía de una latitud a otra en la misma cantidad. Los dibujos de Hevelius, hechos de 1642 a 1644, precisamente cuando se atenuaban ya las últimas manchas antes de la prolongada ausencia, muestran un cambio notable y significativo. El ecuador del sol se había acelerado de tal modo que, para completar una rotación, ¡necesitaba un día menos que en 1625! En efecto, la velocidad de rotación del ecuador, comparada con la velocidad a mayores latitudes solares, se aceleró en



LOS CAMBIOS EN EL CLIMA del mundo coinciden con variaciones de largo periodo en la actividad solar. En la figura se muestra el ciclo de las manchas solares desde 1610 a 1950 *(gris)*, junto con las observaciones orientales (hasta 1750) de grandes manchas observadas a simple vista *(puntos)*. La abundancia relativa del carbono-14 en la atmósfera, determinada a partir del análisis de los anillos de crecimiento anual de árboles, está trazada desde 1050 a 1900 d.C. *(negro)*. La razón del carbono-14 al isótopo común del carbono varía con el nivel de la actividad solar; se produce más carbono-14 en la atmósfera cuando el sol está en calma que cuando está en actividad. Aquí, para poder comparar directamente con la curva de la actividad solar, se invierte la curva de la abundancia del carbono-14, de modo que la abundancia

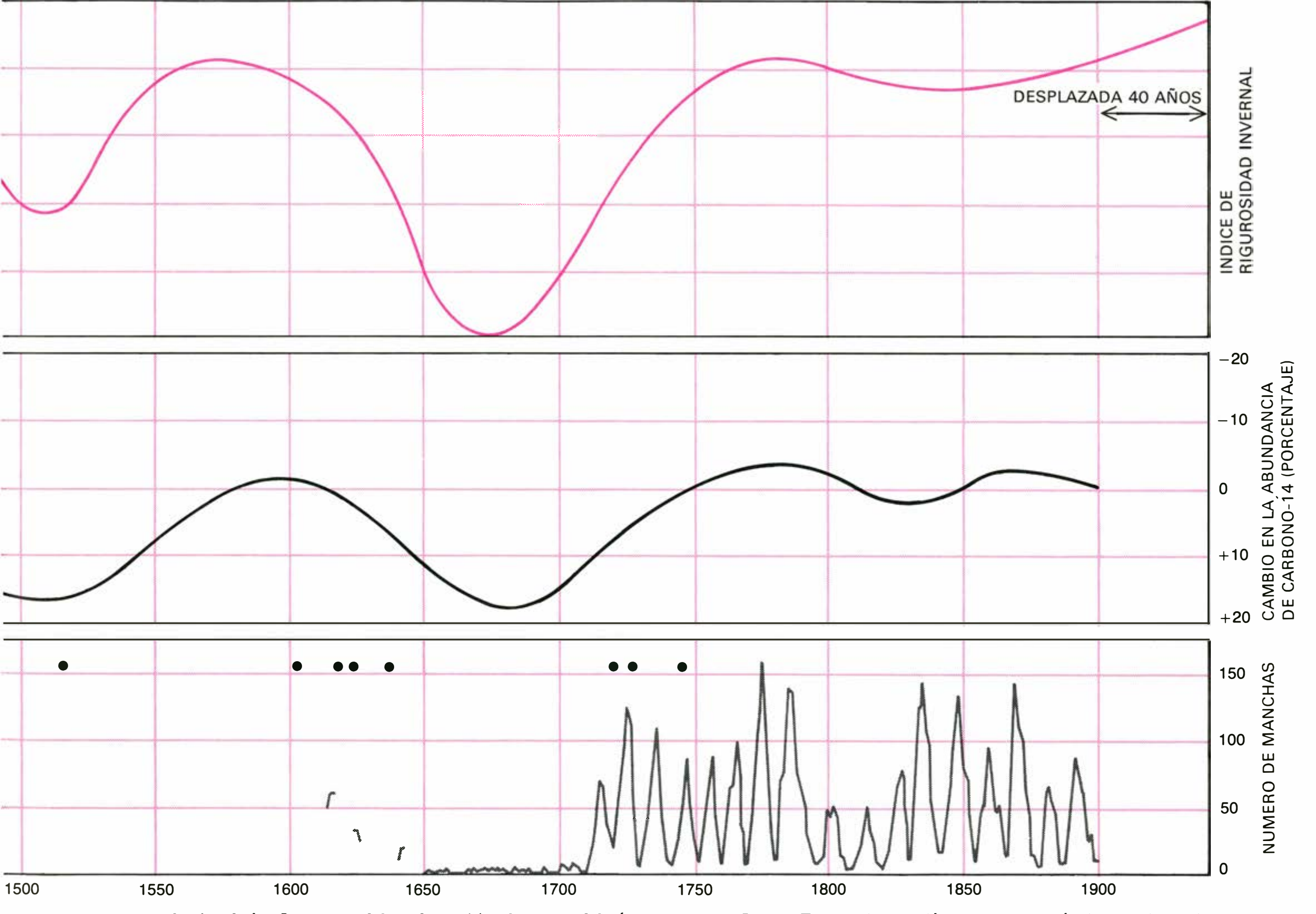
un factor de tres. Al detectar este cambio en el modelo de la rotación de la superficie solar en el preciso momento en que las manchas empezaban a desaparecer ¿habíamos sorprendido al que perpetró el crimen? ¿Fue el cambio en la rotación diferencial del sol causa o efecto del mínimo de Maunder?

Al llegar a este punto de mi rastreo policial, el aspecto del caso había cambiado completamente. Quedaban ahora escasas dudas respecto de que Maunder y los antiguos relatos estuvieran en lo cierto. Faltaba una pista: los anillos de crecimiento anual de árboles viejos.

El isótopo radiactivo carbono-14 se forma en la alta atmósfera de la tierra por la acción de los rayos cósmicos galácticos. El flujo de estos rayos cósmicos está modulado a su vez por la actividad solar, que altera el extenso campo magnético del sol. Cuando éste se halla en gran actividad, su extenso campo magnético protege a la tierra de algunos de los rayos cósmicos galácticos; la cantidad de los que chocan con la alta atmósfera es menor y se forma menos carbono-14. Cuando el sol se halla en un estado de baja actividad, su extenso campo magnético se debilita; la tierra recibe una dosis mayor de rayos cósmicos y aumenta la cantidad de carbono-14 en la atmósfera. Es verdad que, sol aparte, hay otras influencias que condicionan la producción de carbono-14, pero si tuviéramos un registro o historial exacto de la cantidad de carbono-14 que había en la atmósfera en el pasado, tendríamos también un indicio de la actividad del sol en el pasado.

Los árboles proporcionan ese registro, nítidamente separado en divisiones de un año por los anillos anuales de crecimiento. El carbono-14 formado en la alta atmósfera termina por penetrar en los árboles a través de la asimilación del dióxido de carbono en la fotosíntesis. Analizando la madera de cada anillo se puede medir la relación entre carbono-14 y carbono normal que existía en el dióxido de carbono en el momento de formarse dicho anillo. Este laborioso análisis se ha efectuado durante muchos años en varios laboratorios de anillos de árbol, principalmente porque se necesita una secuencia de la abundancia del carbono-14 hasta nuestros días para una datación exacta mediante ese isótopo en arqueología y otras disciplinas. Varios especialistas en este trabajo han señalado el valor potencial del registro del carbono-14 como indicador de la actividad solar en el pasado.

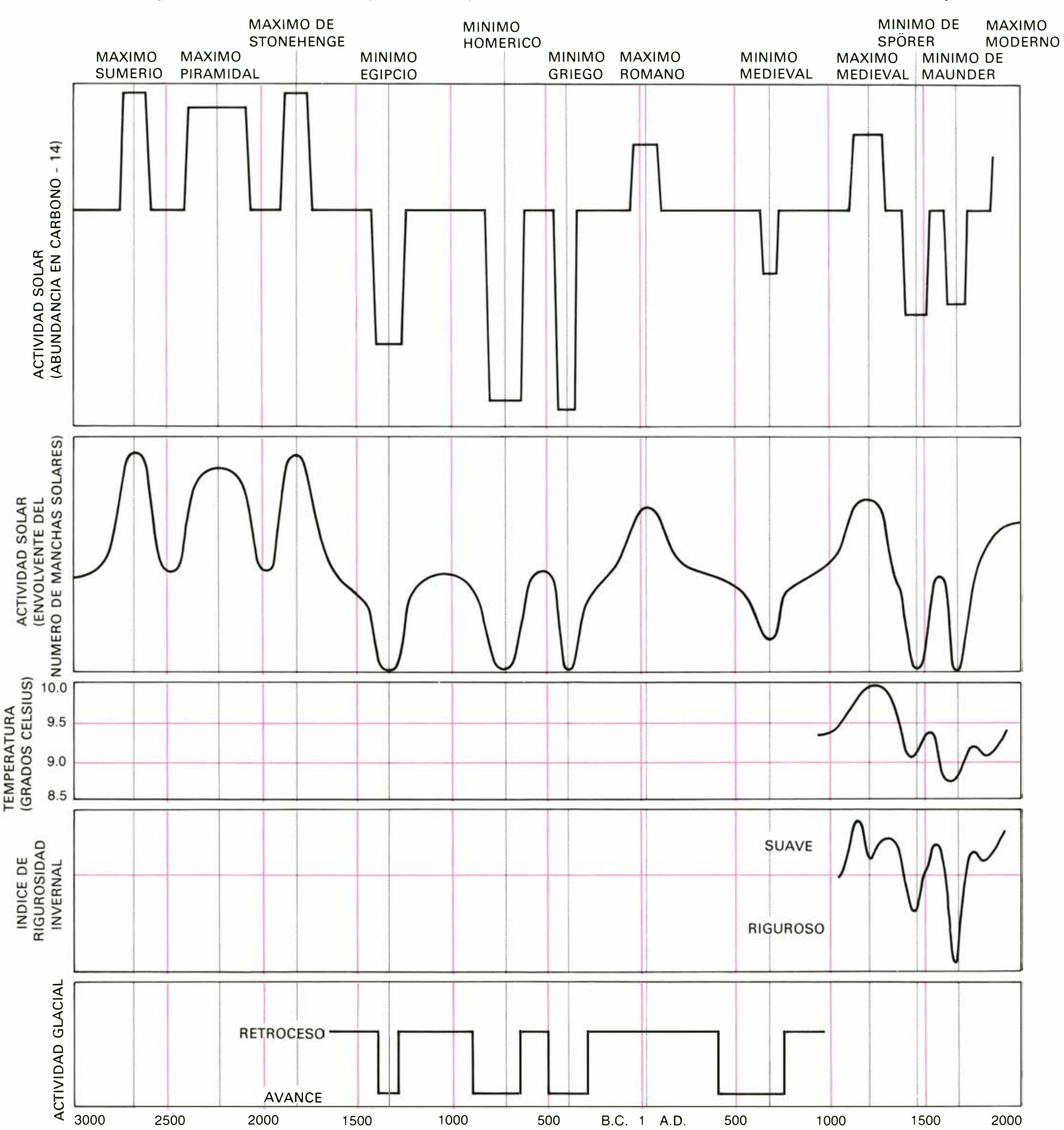
Ya en 1958 el investigador holandés Hessel De Vries llamaba la atención sobre una sorprendente anomalía en la abundancia del carbono-14 en los anillos de árbol de la segunda mitad del siglo XVII y de principios del XVIII. En esos años aumentó bruscamente la abundancia de carbono-14 exactamente como



aumenta hacia abajo. La curva del carbono-14 y la curva del número general de manchas solares observadas se ajustan bastante bien. Los máximos y los mínimos en el nivel de la actividad solar coinciden también con los máximos y los mínimos de la curva que representa el índice de rigurosidad de los inviernos de París y Londres *(color)*, determinados a partir de reseñas históricas por el climatólogo británico Hubert H. Lamb. En esta ilustración, la curva del índice de rigurosidad de los inviernos está desplazada para compensar el hecho que hay un decalaje de unos 40 años entre la producción de carbono-14 en la alta atmósfera y el ingreso del carbono-14 en los árboles. Después del 1900, en que la actividad solar y la temperatura de la tierra aumentaron, el registro del carbono-14 no es válido para representar la actividad solar.

debía ocurrir si el sol hubiera estado desacostumbradamente inactivo durante dicho periodo. El efecto De Vries ha sido confirmado con datos de anillos de árbol de todo el mundo. No obstante, su asociación con una correspondiente anomalía en el comportamiento del sol no se estimó en todo su valor, principalmente por los escasos conocimientos que había sobre la historia de la actividad solar. Ahora parece seguro que la notoria desviación que se observa en el registro del carbono-14, establecida por primera vez por De Vries, confirma enérgicamente la existencia del mínimo de Maunder.

La verificación histórica del mínimo de Maunder ha hecho posible de este modo un apasionante avance en la recuperación de la historia del sol. A través del análisis de los anillos del pino setáceo



LAS VARIACIONES DEL SOL DESDE LA EDAD DE BRONCE se pueden deducir de la abundancia de carbono-14 en los anillos de crecimiento del pino setáceo *(Pinus aristata)*, **usando el mínimo de Maunder para calibrar el diagrama del carbono-14 en función del cambio solar. En el diagrama del carbono-14 aparecen oscilaciones muy pronunciadas representadas por una línea quebrada** *(arriba)*, **cuyas alturas y duraciones se dedujeron del registro del carbono-14. Debajo de esta curva, el autor ha deducido la curva histórica suavizada** *(segunda desde arriba)* **de la actividad solar, que puede interpretarse como una envolvente de largo periodo de la amplitud de un posible ciclo de manchas solares. A partir de ambas curvas se deduce evidentemente que durante 5000 años ha habido, por lo menos, 12 oscilaciones solares tan sobresalientes como el mínimo de Maunder; los nombres sugeridos para variaciones anteriores se han asignado según la época histórica. Se muestra un cálculo de la temperatura media anual en Inglaterra desde alrededor del año 1000 d.C.** *(centro)* **y también la curva de la rigurosidad invernal de París y Londres** *(segunda desde abajo)*. **La línea negra en escalera, al final del gráfico, muestra las épocas en que los glaciares alpinos avanzaban o retrocedían. Durante 5000 años, todas las curvas climatológicas parecen elevarse y descender de acuerdo con la actividad solar.**

(Pinus aristata), el ser viviente más viejo de la tierra, el registro actualmente disponible del carbono-14 abarca por lo menos hasta 5000 años antes de Cristo. El aspecto dominante de este extenso registro es una lenta modulación de la abundancia del carbono-14 que pone de manifiesto un aparente periodo de unos 10.000 años. Esta modulación es inducida por cambios en la intensidad del campo magnético de la tierra, el cual actúa como otra protección contra los rayos cósmicos galácticos. En el registro del carbono-14 también están presentes otras muchas oscilaciones de periodo más corto. Ahora bien, yo había establecido que la más reciente de estas desviaciones de corto periodo del carbono-14 correspondía a una anomalía solar que parecía bien establecida por pruebas históricas directas e indirectas. Aunque esta posible asociación había sido señalada antes, por primera vez supe con precisión qué clase de cambio en el sol producía una desviación de un determinado valor en el registro del carbono-14. Al fin podíamos calibrar el registro del carbono-14 en función del cambio solar. Con el mínimo de Maunder como clave podíamos descifrar la historia largamente escondida del sol, no sólo hasta la época de Luis XIV o de Galileo, sino más atrás, hasta antes del comienzo de la edad del bronce, hace unos 7000 años.

La solución de un caso había desenmascarado otro caso más complejo y posiblemente más importante.

Por lo menos 12 aspectos tan sobresalientes como el mínimo de Maunder resaltan en el diagrama del carbono-14 de los últimos 5000 años. Cada uno dura entre 50 años y varios siglos. Algunos apuntan a una actividad solar muy reducida, como el mínimo de Maunder. Uno de éstos es el mínimo de Spörer, que abarca desde alrededor del 1400 al 1510 d. C., época en que se tiene otra vez poca información de auroras y manchas observadas a simple vista y en que no hay referencias acerca de una estructura coronal, procedentes de algún observador de un eclipse solar total. Otras variaciones mayores apuntan en dirección opuesta; por consiguiente, corresponden a eras de actividad solar insólitamente alta, tal vez más alta incluso que la que hemos conocido en tiempos modernos. Una de estas variaciones, que llamo el máximo medieval, cae entre el 1000 y el 1300, aproximadamente. Otra, según creo, es la que vivimos ahora, y se caracteriza por la elevación del nivel de actividad solar que siguió al mínimo de Maunder y que puede o no haber alcanzado su cúspide en el máximo de manchas solares de 1959, el más alto registrado hasta la fecha. Sólo durante este máximo moderno hallamos información sobre la estructura de la corona y sobre la elevada frecuencia de auroras.

No hay seguridad al interpretar la era presente de actividad solar a partir de las pruebas referentes al carbono-14 a causa del efecto Suess: existe una marcada dilución de la abundancia del carbono-14 en los anillos de los árboles desde finales del siglo XIX. Se ha atribuido a la introducción de grandes cantidades de carbono con escaso contenido de carbono-14 en la atmósfera a través del incremento de la combustión de combustibles fósiles, en todo el mundo, durante el siglo pasado. Al parecer, el efecto Suess ha dejado caer un telón sobre lo que podemos enterarnos, a partir de los cambios naturales inducidos en la abundancia relativa de carbono-14 en la era moderna, impidiéndonos emplear el estado actual de la actividad solar para calibrar cambios pasados en el registro del carbono-14.

Una última comparación queda por hacer, una comparación que, como sugirió una vez Maunder, puede enlazar los cambios solares de largo período con importantes efectos en la tierra. El mínimo de Maunder se corresponde casi exactamente con la más fría de las oscilaciones de la "pequeña glaciación", un periodo de frío poco común en Europa, que abarcó desde el siglo XVI hasta principios del siglo XIX. En los extremos más fríos de ese periodo, la temperatura media fue de un grado Celsius más baja que ahora, de acuerdo con el climatólogo británico Hubert H. Lamb. Los glaciares alpinos avanzaron más de lo que habían avanzado desde la última glaciación mayor, hace 15.000 años. En este periodo también la colonia escandinava establecida en el sudoeste de Groenlandia desapareció por completo, separada del resto del mundo por hielo flotante que año tras año seguía sin deshelarse. ¿Podía relacionarse esta anomalía en el clima –el más riguroso de los últimos milenios– con la larga ausencia de manchas solares? La desaparición de las manchas solares y la distorsionada configuración de la rotación solar, ¿serían acaso indicios de que la energía de la radiación solar hubiese disminuido ligeramente? Los modelos modernos del clima muestran que los periodos de frío global tan gélidos como la "pequeña glaciación" pueden producirse por una disminución de no más del 1 por ciento en el total de la radiación solar, un cambio sutil que si se prolongase por unas décadas sería apenas detectable mediante mediciones directas.

Por otra parte, la coincidencia podría ser simplemente que el mínimo de Maunder y la pequeña glaciación fueran anomalías no relacionadas entre sí. Un colega me advirtió una vez del riesgo que corría al hacer una asociación tan simple; me indicó que igualmente podía postularse una relación entre el mínimo de Maunder y el reinado de Luis XIV que coincidió con él. ¿Podríamos decir que un prolongado mínimo de manchas solares produce un Rey Sol?

Ahora que el documento más amplio de la historia solar está a nuestro alcance en la forma de carbono-14, podemos analizar una conexión significativa entre el cambio solar y el clima, comparando cada una de las oscilaciones solares importantes que aparecen en el diagrama del carbono-14 con la historia del clima en la época correspondiente. Al hacer esta comparación estaremos muy limitados por la incertidumbre en el registro del clima; hoy nos es dado conocer la historia del sol ¡mejor que la historia de nuestro propio planeta!

He comparado la historia del sol, testimoniada por el carbono-14 con la historia del clima del mundo deducida, por Lamb y otros, de los archivos históricos y del avance y del retroceso de los glaciares alpinos. El ajuste es casi como el de una llave en su cerradura. Cada disminución de la actividad solar, verbigracia, el mínimo de manchas solares de Maunder, se corresponde con una época de avance de glaciares en Europa; cada aumento de la actividad solar, por ejemplo, el máximo medieval, se corresponde con una época de retroceso de los glaciares. El mínimo de manchas solares de Spörer se corresponde en profundidad y duración con el primer mínimo de temperatura de la pequeña glaciación. El máximo solar medieval se corresponde con su época cálida en que la temperatura media del mundo fue tanto o más elevada que la actual. Estos primeros resultados de la comparación de la historia solar con el clima hacen que parezca que los cambios solares sean el agente dominante de los cambios del clima que duraron entre 50 años y varios siglos.

Curiosamente, el aparente vínculo entre cambio solar y cambio climático poco puede revelar acerca de si los cambios solares de corto periodo, tales como el ciclo de 11 años de manchas solares, tienen algún efecto importante sobre la evolución climática a corto plazo. En el mínimo de Maunder o en el máximo medieval no hemos estudiado cada subida y bajada del ciclo de las manchas solares, sino la envolvente de largo periodo que enlaza los extremos de muchos ciclos. Me parece posible que la envol-

vente lentamente variable quizá refleje suaves cambios de algunas unidades porcentuales en la energía total irradiada por el sol, cambios que podrían ser bastante independientes de que el ciclo de 11 años esté en su máximo o en su mínimo. Una energía solar variable podría modular la amplitud o intensidad de una sucesión continua de ciclos undecenales de manchas solares, cada uno de los cuales seguiría presentando una cúspide y un valle. Como en una señal radioeléctrica de amplitud modulada, el mensaje sería transportado no por los ciclos aislados de la onda continua, sino por los de amplitud variable, que es más reconocible en la envolvente de las cúspides.

La intensidad del ciclo de las manchas solares podría ser modulada por la dinamo solar, cuando cambios lentos en el flujo de energía solar alteran la estructura de la zona convectiva del sol y, por consiguiente, el modelo de circulación en su superficie. La cantidad de radiación emitida por el sol podría ser casi enteramente independiente de la fase del ciclo undecenal de manchas solares. Tal mecanismo podría explicar por qué en las investigaciones acerca de la conexión entre el sol y el tiempo atmosférico han surgido tantos contratiempos cuando se ha tratado de hallar correlaciones con el ciclo undecenal de las manchas solares, que podría ser sólo la frecuencia de la onda portadora.

Parece pues que Maunder y Spörer estuvieron acertados y que la mayoría restante de nosotros ha estado equivocada. Como ocurre muchas veces en la embestida de la ciencia moderna, habíamos olvidado el pasado con excesiva precipitación, habíamos olvidado la no muy perfecta genealogía del ciclo de las manchas solares y el hecho de que también ese ciclo en otro tiempo apareció de improviso. Habíamos adoptado una especie de uniformitarismo solar, afirmando que el comportamiento moderno del sol representaba su comportamiento normal durante un lapso mucho más largo.

Como personas y como científicos, siempre quisimos que el sol sea mejor que las otras estrellas y mejor que lo que realmente es. Hace mucho tiempo sostuvimos que era perfecto, y cuando el telescopio mostró que tenía manchas nos conformamos pensando que por lo menos era regular en su comportamiento. Ahora parece que no es ni lo uno ni lo otro y que además resulta probablemente inconstante, pero el saberlo abre camino a un conocimiento mucho más completo del sol y de sus importantes influencias sobre la tierra.

La investigación del herbario

Se puede obtener una valiosa información sobre plantas poco conocidas, capaces de suministrar drogas nuevas, a través del examen de las etiquetas que acompañan a los millones de ejemplares existentes en las colecciones botánicas

Siri von Reis Altschul

En vista de que puede que existan alrededor de 800.000 especies de plantas en el mundo, resulta curioso considerar cuán pocas han sido explotadas por el hombre. Las tres especies de cereales más importantes y quizás otras diez más, cultivadas ampliamente, son lo que mantiene a la población humana del mundo alejada del hambre; y sólo un puñado de drogas vegetales ha servido a la civilización occidental durante varios miles de años. Surge una pregunta inmediata: ¿no estaremos desperdiciando algún recurso? Es indudable que sí. Entre las 250.000 o más especies conocidas, nombradas y descritas, hay otras plantas que podrían cultivarse para mejorar el suministro mundial de alimentos o cuyos compuestos químicos, biológicamente activos, podrían aislarse con fines medicinales. ¿Qué hacer para encontrarlas?

Años atrás, se me ocurrió que los herbarios de todo el mundo constituían una fuente de información accesible e inmensamente rica sobre plantas de posible valor alimentario o medicinal. Hay alrededor de 1800 herbarios públicos en el mundo, que albergan más de 175 millones de ejemplares de plantas secas y prensadas y que representan aproximadamente 250.000 especies diferentes conocidas y algunas otras no identificadas todavía. Había visto suficientes de tales colecciones como para saber que, en algunos casos, las notas de campo unidas a los ejemplares encierran información etnobotánica: datos sobre los conocimientos locales concernientes a la planta y sobre los usos dados a la misma por los pueblos indígenas de la región. Me di cuenta de que para los taxónomos y estudiosos de la evolución de las plantas, que son quienes tradicionalmente trabajan en los herbarios, tal información pudiera resultar muchas veces periférica e incidental, por lo que la mayor parte de ellos la habrían pasado por alto. Decidí continuar con el tema. Una exploración de cinco años de las colecciones de Gray Herbarium y Arnold Arboretum, en la Universidad de Harvard, me convenció de que en las etiquetas de los herbarios existen valiosos datos, recogidos y conservados desde hace varios cientos de años, procedentes de todas las partes del mundo, que esperan ser sacados a la luz.

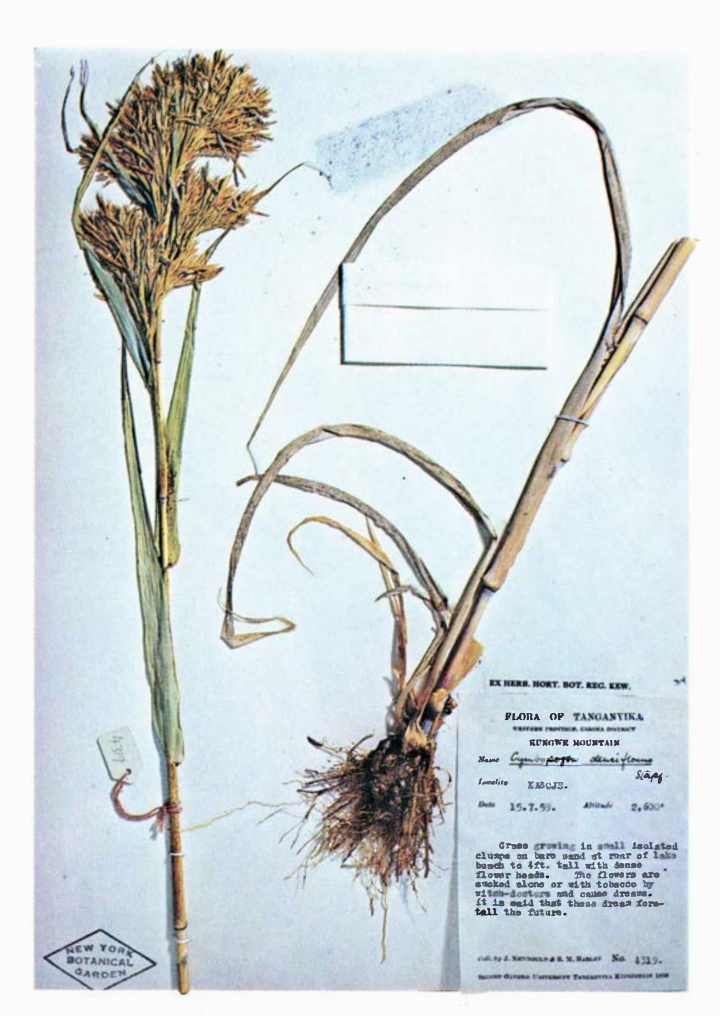
No resulta sorprendente, en absoluto, que los herbarios sean depositarios de orientaciones sobre los usos de las plantas poco conocidas. Desde sus comienzos, el estudio de las plantas estuvo profundamente relacionado con sus usos, y, en particular, con sus usos como drogas. De hecho, la botánica y la medicina han estado íntimamente entrelazadas hasta tiempos muy recientes. Los pueblos primitivos de todas las épocas tuvieron amplios conocimientos de la vegetación de su entorno. Esto es igualmente cierto de los pueblos primitivos actuales, cuyas taxonomías populares están basadas fundamentalmente en consideraciones utilitarias, tal como lo estuvieron las de las civilizaciones primitivas. Unos 5000 años a. C., los chinos habían desarrollado ya importantes farmacopeas vegetales, es decir, series de drogas aceptadas y derivadas de las plantas. Lo mismo sucedió entre los babilonios, los asirios y, más tarde, los hebreos. Los trabajos de Aristóteles e Hipócrates revelan que los griegos conocían muchas de las drogas que nosotros usamos corrientemente. Plinio y Galeno escribieron sobre drogas vegetales, y Dioscórides, contemporáneo de Plinio, escribió la gran *De materia medica*, que se ocupa de todas las sustancias medicinales conocidas en su tiempo, incluyendo unas 600 plantas, y que ha sido el tratado clásico sobre el tema durante quince siglos.

La invención de la imprenta hacia 1440 preludió la época de los herbolarios, compendios ilustrados de datos (algunos verdaderos y otros falsos) sobre las plantas, con especial énfasis en su valor medicinal o alimentario. Los herbolarios estimularon un nuevo interés por la botánica, tanto de las plantas que crecían al norte de los Alpes como de las que habían sido conocidas por el mundo antiguo. Debido a que ahora la palabra escrita y los dibujos podían duplicarse fácilmente y distribuirse con amplitud, resultaba posible comparar las plantas encontradas en diferentes lugares. En consecuencia, se incrementaron los esfuerzos para clasificar las múltiples especies. No obstante, no parece haber habido intento alguno de crear colecciones sistematizadas de plantas secas y prensadas hasta mediados del siglo XVI, cuando el botánico italiano Andrea Cesalpino y sus colaboradores comenzaron a conservar ejemplares para futura referencia y autentificación. Algunos botánicos centroeuropeos siguieron el ejemplo. Pronto se apreciaron claramente las ventajas de mantener colecciones de ejemplares conservados y surgieron los herbarios en la forma en que los conocemos actualmente.

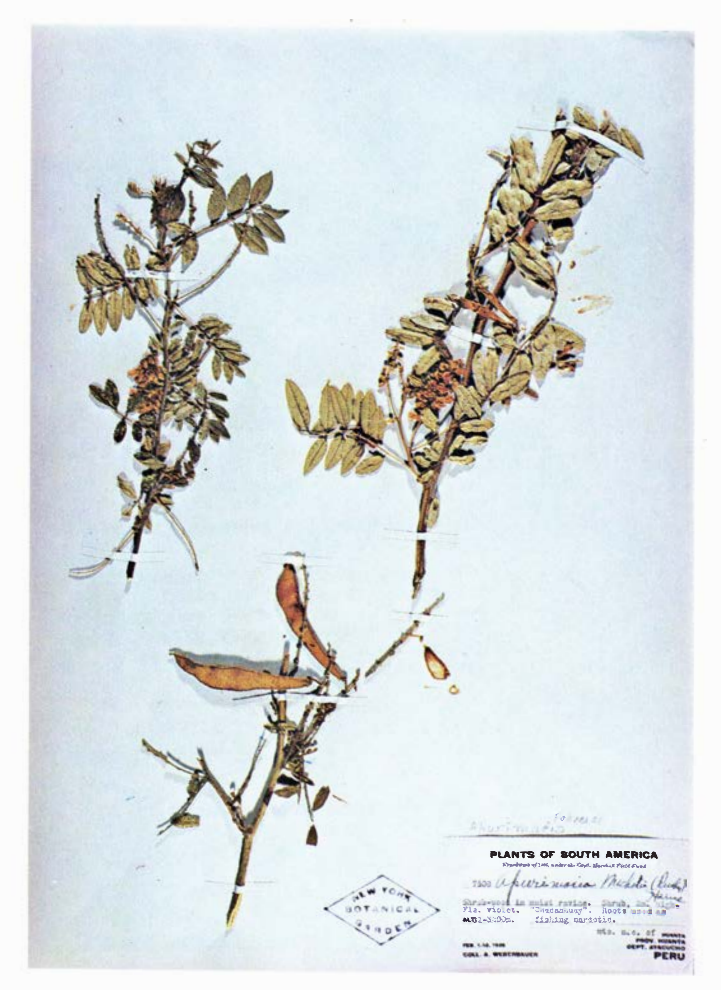
Herbolarios, herbarios, el cambiante clima intelectual de Europa y el descubrimiento del Nuevo Mundo con su enorme riqueza natural, todo contribuyó

PLIEGO DE HERBARIO que presenta un ejemplar de planta seca con las notas de campo detalladas; éstas incluyen datos etnobotánicos que sugieren usos alimentarios o medicinales. El ejemplar en cuestión es un bambú del género *Guadua*, de la familia de las gramíneas, recolectada en la antigua Guayana inglesa. En la nota se lee: "se dice que es venenosa. Usada en las puntas de flechas".


Bambuseae 34
FOREST DEPARTMENT OF BRITISH GUIANA.
Date. 17/10/52
Field No. G 432.
Record No 7447
Alt.
Loc. Near Ghodikar landing 10' from River.
Desc. Woody grass to 30', scrambling or erect. Stems to 3" thick, very small hole. Twig nodes at 3" with tufts of Compound leafy shoots. The leaves lanceolate, 4" long, pale green with sheathing bases.
Said to be poisonous. Used in arrow points.
Vern. Name Baundoo Raffu
Bot. Name Guadua ?
Family GRAMINEAE.
FORESTRY DEPT.—No. 5
NEW YORK BOTANICAL GARDEN


EX HERB. HORT. BOT. REG. KEW.
FLORA OF TANGANYIKA
KUNGWE MOUNTAIN
Name
Locality KABOJE.
Date 15.7.59. Altitude 2,600'
Grass growing in small isolated clumps on bare sand at rear of lake beach to 4ft. tall with dense flower heads. The flowers are smoked alone or with tobacco by witch-doctors and cause dreams. It is said that these dreams foretell the future.
No. 4319.
NEW YORK BOTANICAL GARDEN


FLORA DE VENEZUELA
No. 99613
"cobalonga"
Tree 15 m.; leaves chartaceous, deep green above, paler green below; fruit rusty brown. Seed when dried placed in infusion, used as drink to help mothers in child birth
SIERRA DE PERIJA
NEW YORK BOTANICAL GARDEN


PLANTS OF SOUTH AMERICA
Fls. violet.
fishing narcotic.
PERU
NEW YORK BOTANICAL GARDEN


FIELD MUSEUM OF NATURAL HISTORY
November 22, 1966
NEW YORK BOTANICAL GARDEN

a un cambio de mentalidad, en virtud del cual, aunque la medicina conservaba su interés central por las drogas derivadas de las plantas, los botánicos sintieron cada vez más fascinación por las plantas mismas. Pronto se puso de relieve que los diferentes sistemas por los que se las clasificaba eran embarazosos e inadecuados para la pujante tarea. Sobre tal confusión impuso un orden el gran botánico y zoólogo sueco Carl von Linné, o Linnaeus, en el siglo XVIII. Inventó un sistema que, aunque era artificial, resultaba práctico para clasificar las plantas y se basaba fundamentalmente en las características estables y ostensibles de la flor: el número, posición o longitud relativa de los estambres. Y lo que es más importante: estableció también un método uniforme para designar las especies por medio de dos nombres latinos, método que pudo ver aceptado internacionalmente en vida. El mismo Linneo dio nombre a más de 10.000 especies. Su *Species plantarum* de 1753, en el que describe y da nombre a todas las especies conocidas, y su *Genera plantarum*, en el que describe todos los géneros conocidos, siguen siendo obras clásicas a las que los botánicos taxónomos acuden regularmente. Es notable que libros escritos hace dos siglos permanezcan como piedra angular de una disciplina contemporánea.

Linneo vivió en una época de gran preocupación por el saber enciclopédico, sostenida por la fe en la existencia de un mundo natural fijo, que fue dotado de un orden inmutable en un único instante de creación. Parecía que la observación cuidadosa de todas las partes de aquel mundo revelaría con toda seguridad su coherencia subyacente. Con esta inspiración, la exploración botánica fue proseguida con vivo fervor, principalmente por un notable grupo de "apóstoles" de Linneo, como él los llamaba, los cuales emprendieron numerosas expediciones a las Américas, a los mares del Sur y al Extremo Oriente. Continuaron otros investigadores, y los ejemplares fueron depositados en los herbarios de Europa. Gracias al sistema de Linneo, estas masas de plantas pudieron ser clasificadas y archivadas convenientemente, pudieron describirse sucintamente nuevas especies y pudo divulgarse su descripción.

En el siglo XIX, las colecciones botánicas recibieron una nueva clase de impulso, cuando la mente del hombre occidental empezó a abrirse hacia nuevas ideas sobre sí mismo y sobre sus relaciones con el resto del mundo natural. Conforme la idea de evolución se fue adentrando en la conciencia humana se puso en marcha una nueva ola de exploraciones, ahora en busca de un continuo cambio en la naturaleza.

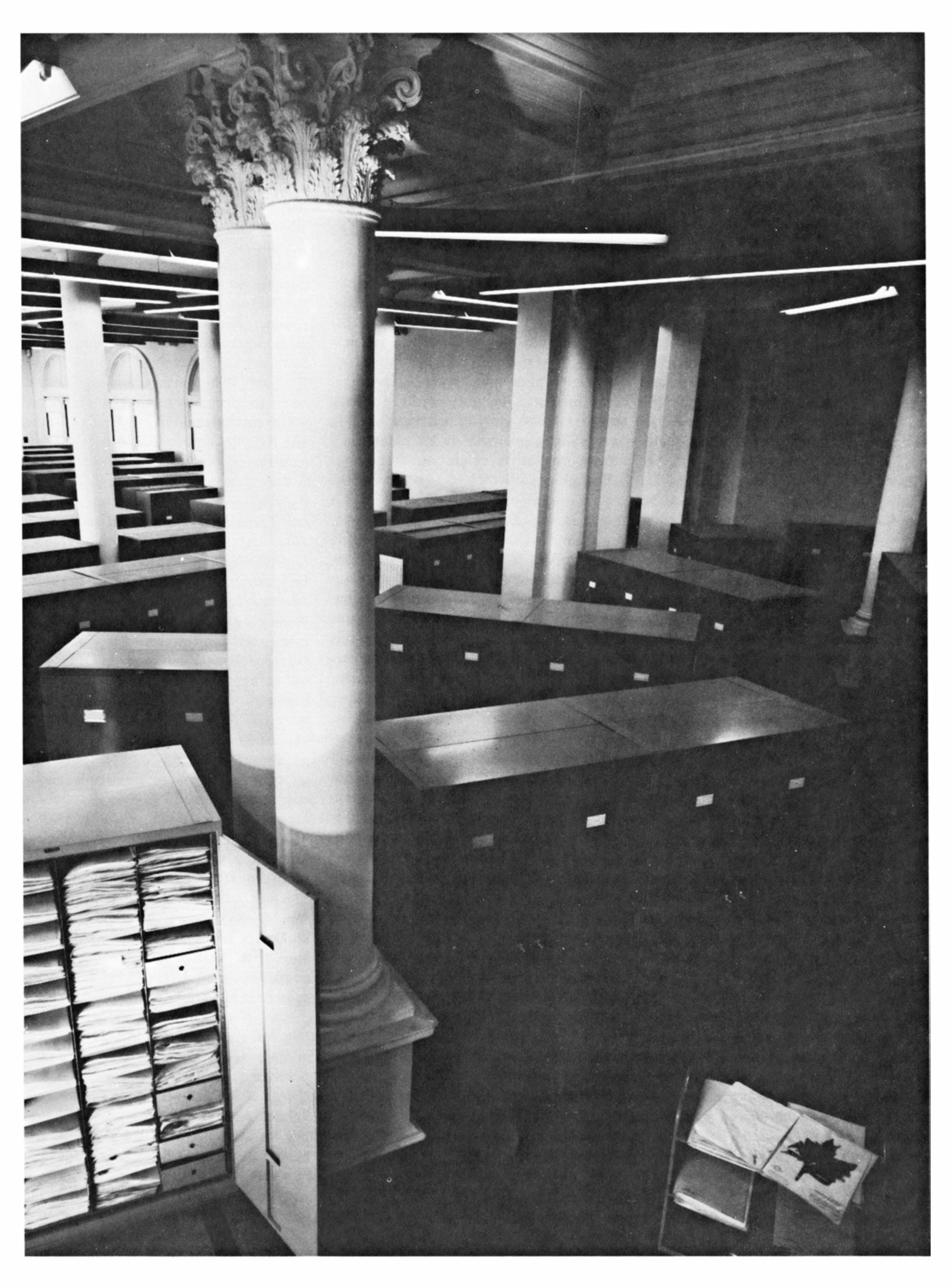
Los científicos, pues, anhelaban encontrar no sólo qué especies había, sino también cómo se habían diferenciado, y conocer qué grupos de plantas se hallaban relacionadas y cómo. Se recolectaron y archivaron los ejemplares en herbarios para una posterior investigación mediante nuevas técnicas, que iban desde la citología y la fisiología hasta la geografía y ecología de las plantas. La lógica evolutiva exigía modificar muchos grupos artificiales que se habían creado según otras consideraciones. La clasificación de Linneo, basada fundamentalmente en los estambres, fue una de las primeras víctimas. Las nuevas investigaciones requirieron extensos estudios de herbario y de laboratorio, generalmente acompañados por ulteriores trabajos de campo y recolecciones más minuciosas. La herborización se convirtió en el rasgo principal de la botánica sistemática de finales del siglo XIX, que combinaba la taxonomía (la clasificación, como expresión de afinidad evolutiva) y la nomenclatura. Realmente, la sistemática se convirtió casi en una obsesión en los departamentos de botánica de las universidades de América y Europa, hasta el punto de que llegó a producirse una reacción contra el coleccionismo, la clasificación y la herborización. Algunas colecciones fueron eliminadas y deshechas. Hoy la recolección de nuevos ejemplares y su conservación en herbarios se consideran como fundamentales para estudios básicos de biología y evolución de las plantas, que son los que ocupan en gran parte a los botánicos. Y la sistemática continúa desempeñando un papel exclusivo: recoge e integra los conocimientos de un creciente número de disciplinas, al tiempo que las sirve al identificar nuevas especies y esclarecer las posibles relaciones dentro del reino vegetal.

USOS MEDICINALES sugeridos por las notas de los pliegos que se encuentran en la página opuesta. *Cymbopogon densiflorus (arriba, izquierda),* **una hierba de Tanganica (actual Tanzania), tiene flores que "los hechiceros se fuman solas o con tabaco y producen un estado hipnótico. Se dice que en este trance se adivina el futuro".** *Billa columbiana (arriba, derecha),* **de Venezuela, tiene un fruto "pardo herrumbroso" cuyas semillas secas "se ponen en infusión y se usan como bebida para facilitar el parto". Las raíces de** *Apurimacia michelii (abajo, izquierda),* **una legumbre arbustiva de Perú, "se emplean como narcótico de pesca".** *Eleutherine bulbosa (abajo, derecha),* **del Perú lleva la anotación: "Anticonceptivo. Los indios Shipibas la toman como infusión".**

En todo el mundo hay algunos miles de plantas que son apreciadas por sus atributos medicinales, la mayor parte de ellas por los pueblos primitivos o por los curanderos. La moderna farmacología occidental reconoció que muchas de las drogas vegetales tradicionales contienen, en sus tejidos, compuestos particulares que producen efectos específicos sobre el cuerpo humano. Existe un amplio espectro de estos compuestos químicos de actividad biológica, que pueden ser clasificados como alcaloides, glucósidos, aceites esenciales, grasas, resinas, mucílagos, taninos, gomas y otras sustancias. Se estima que más de la mitad de las recetas extendidas por los facultativos en la actualidad contienen una droga vegetal: una droga que o bien procede realmente de una planta o que ha sido sintetizada para duplicar (o para mejorar) el compuesto vegetal.

Por ejemplo, el cólquico, *Colchicum autumnale,* nos suministra el alcaloide colchicina, que es el específico contra la gota. La raíz de ipecacuana, *Cephaelis ipecacuanha,* era un remedio antiguo de los indios sudamericanos cuando fue descubierta en el siglo XVIII por el médico de corsarios Thomas Dover; la raíz de ipecacuana se convirtió en el principal ingrediente de los polvos de Dover y se utiliza hoy día para el tratamiento de la disentería amebiana. La rauwolfia o serpentaria, *Rauwolfia serpentina,* se vendía hace 3000 años en los bazares de la India para las mordeduras de serpiente y como un calmante para los trastornos psíquicos; desde 1940, su principio activo, la reserpina, ha sido reconocido como una droga efectiva, como un tranquilizante primero y ahora sobre todo como un agente para bajar la presión sanguínea. El curare es un veneno de flechas, de rápida acción, utilizado todavía por los indios sudamericanos y obtenido de una mezcla de hasta 30 especies diferentes de plantas. Uno de sus alcaloides activos es la base de varias drogas que tienen importancia en la cirugía moderna como relajantes musculares. La quinina, una droga antimalárica muy usada y barata, procede de la corteza de una especie de *Cinchona,* un arbusto de la misma familia del cafeto (rubiáceas) ya conocido por los incas antes de la conquista española. Los jesuitas llegaron a conocerla bien (de ahí que uno de sus nombres vulgares sea "cáscara de los jesuitas") y la introdujeron en la medicina europea.

Una de las más antiguas drogas vegetales que utilizamos todavía con profusión es la *Ephedra,* un arbusto conocido en China desde hace más de 5000 años;



HERBARIO del jardín botánico de Nueva York. Contiene unos 4 millones de ejemplares, ordenados filogenéticamente por familias y alfabéticamente por géneros y especies, en 1250 cajas de acero. (Los capiteles de las columnas corintias modificadas remedan la hoja de *Acanthus spinosus.*)

su alcaloide, la efedrina, se utiliza hoy como descongestionante nasal y estimulante del sistema nervioso central. La belladona; *Atropa belladonna*, es otra antigua planta medicinal, de la que se extrae un poderoso alcaloide, la atropina, que dilata la pupila de los ojos. La cocaína, que hoy en día suele ser sintética, fue descubierta en el *Erythroxylon coca*, un arbusto cuyas hojas mastican todavía los indios de los Andes y del Amazonas para aliviar el hambre y las fatigas y como un narcótico estimulante. Las hojas de *Digitalis purpurea*, la digital púrpura, contiene un glucósido con efectos terapéuticos específicos sobre el músculo cardíaco, que fue reconocido por la medicina popular ya antes de que el facultativo inglés William Withering demostrase el valor de la planta hacia finales del siglo XVIII; la digital sigue siendo una droga fundamental para el tratamiento de las enfermedades cardíacas.

El cornezuelo del centeno, el hongo *Claviceps purpurea*, que infecta el centeno y otras gramíneas, posee una sustancia venenosa. En el año 944 murieron en Francia 40.000 personas debido a una epidemia de cornezuelo que contaminaba el pan, y tales plagas se repitieron hasta que se identificó la causa en el siglo XVII. Las matronas conocían, sin embargo, desde hacía mucho tiempo, que el cornezuelo aceleraba el parto e inhibía la hemorragia subsiguiente. Hoy en día, un derivado de este hongo parásito constituye el remedio más eficaz, hoy, para los dolores de cabeza; los alcaloides derivados del cornezuelo de centeno se administran para contraer los músculos uterinos en caso de partos difíciles. Desde tiempo inmemorial, los japoneses comían un alga roja llamada *Digenea simplex* para desembarazarse de los parásitos intestinales. El principio activo se encontró hace poco y se trata del ácido kaínico, una sustancia potente que se ha convertido, en la actualidad, en tratamiento corriente contra los parásitos intestinales en el Extremo Oriente.

Todas las culturas humanas tienen sus farmacopeas, y se ha estimado que del 25 al 50 por ciento de las drogas de las farmacopeas no occidentales son eficaces empíricamente. Es decir, cerca de la mitad de las drogas de los repertorios médicos nativos pueden curar o proporcionar alivio realmente, aunque no siempre sea por las razones establecidas por la tradición local. Según Richard Evans Schultes, del museo botánico de la Universidad de Harvard, la investigación del reino vegetal en cuanto a las sustancias de valor farmacológico es tan incompleta que, probablemente, se han identificado menos de la mitad de los compuestos producidos por las plantas, para no hablar de su eficacia.

Ahora bien, ¿cómo pueden los investigadores localizar y evaluar las plantas y sus constituyentes? La fuente habitual de orientación y guía es la literatura publicada: trabajos clásicos, primeras descripciones de las plantas y sus efectos e informes posteriores de los exploradores y antropólogos. Sin embargo, incluso si se va a las fuentes originales, resulta que no suelen citarse las muestras, de tal manera que es imposible comprobar la identidad de las plantas aludidas en el texto. Un español del siglo XVI pudo haber hablado de los indios del Nuevo Mundo que recolectaban una "planta con hojas anchas y semillas pequeñas que tomaban por la nariz y que producía la más extraordinaria reacción, en que esos indios caían en un frenesí acompañado de maravillosas visiones", pero nosotros sólo podemos hacer conjeturas sobre la identidad de la planta. Incluso los informes modernos sobre estudios antropológicos de las sociedades primitivas han omitido, generalmente, las necesarias identificaciones botánicas; sólo muy recientemente han comenzado tales informes a citar los ejemplares "comprobantes" o de referencia.

Queda la alternativa de volver a las floras, que son recopiladas por los botánicos y catalogan las especies que crecen en una región determinada, pero tales listas generalmente omiten los datos etnobotánicos. En fin, el investigador de nuevas drogas vegetales puede situarse él mismo sobre el terreno y tratar de descubrir qué plantas se consideran valiosas y para qué usos. Esto lleva mucho tiempo y se requiere gran preparación simplemente para poder entenderse con la gente; no digamos ya para conseguir intimar con los brujos o ancianos en el grado necesario para que compartan sus bien guardados secretos. Sustituyendo el saber local por la química, el investigador de campo puede acudir a ensayos rápidos que revelen la presencia o ausencia de ciertos compuestos químicos en una planta. Desgraciadamente, hay que decidir por adelantado qué clase de compuestos son los que se buscan, realizando entonces los ensayos pertinentes para aquellas sustancias en particular, por ejemplo, alcaloides o glucósidos; no existen experimentos de campo que puedan revelar todos los constituyentes de una planta o cuáles sean sus posibles usos.

La arqueología ha sido también una fuente de información sobre los usos de las plantas. Gracias a ella hemos aprendido mucho sobre el origen y desarrollo de las plantas alimenticias en particular. Hay pocos indicios de plantas medicinales en los yacimientos de los asentamientos primitivos humanos. De cualquier modo, los restos arqueológicos se encuentran generalmente resecados, carbonizados, o conservados en agua y son normalmente fragmentarios y frágiles. El estudio de coprolitos (heces fosilizadas) revela qué clase de alimentos ingería el hombre primitivo, pero no por qué era consumido cada alimento.

Nos quedan, pues, los herbarios, que surgieron fundamentalmente en interés de la medicina y se aprovechan ahora para servir de nuevo a la medicina, y quizá también a la ciencia de la alimentación y de la nutrición. Conforme los herbarios se han ido multiplicando y han ampliado sus colecciones a lo largo de los siglos, los coleccionistas fueron preocupándose por hacer anotaciones de campo sobre los usos y los presuntos efectos de las plantas que recogían. Sin ir más lejos, en el siglo XIX muchos recolectores eran facultativos, incluyendo, por ejemplo, a Asa Gray y John Torrey en los Estados Unidos. Las observaciones de primera mano que ellos hacían las escribían sobre etiquetas que pegaban a los ejemplares, de tal modo que no hay duda sobre la planta a la que se refiere la nota. Debido a que esas notas de campo indican dónde fue recogida la planta, los investigadores pueden volver a la localidad de origen para obtener material en abundancia, si fuese necesario para análisis químicos. La mayoría de dichas notas nunca han sido publicadas.

Fue en el curso de un estudio taxonómico de las colecciones de Harvard, que realicé a finales de los años 50, cuando yo misma comencé a impresionarme por la información contenida en los pliegos del herbario, información generalmente ignorada. Estaba trabajando sobre un raro género narcótico y alucinógeno sudamericano y estaba buscando cualquier tipo de información etnobotánica. La literatura publicada, mucha de la cual se basaba en datos de los días de la conquista española, raramente citaba las muestras de referencia, y, por tanto, tuve una enorme alegría al hallar datos que acompañaban a un ejemplar. ¿En qué medida, me preguntaba, las enormes colecciones de Harvard pueden contener información importante y no publicada sobre medicina popular referente a otras plantas? ¿Merecería la pena examinar cada uno de los dos millones y medio de pliegos de herbario para sacar tal información? Presentía que sería importante sacar el mayor partido posible de los fragmentos de información



CARL VON LINNE (LINNAEUS) elaboró el sistema de nomenclatura que facilitó el registro y la clasificación de las especies. Su primera expedición importante fue a Laponia en 1732; en este retrato aparece posando con traje regional; sostiene una *Linnaea borealis*, llamada así en su honor.

reunidos por cientos de coleccionistas y registrados por ellos con la esperanza de poder contribuir así no sólo a la taxonomía, sino también al avance del conocimiento en general. Una vez terminé mi trabajo doctoral, comencé a investigar con la colaboración de dos ayudantes.

Buscábamos y registrábamos cualquier nota que sugiriese constituyentes de actividad biológica: aquéllos que tuviesen efectos sobre tejidos vivos, desde irritantes de la piel y venenos, pasando por los medicamentosos de cualquier tipo, a los alimenticios, bebidas y especias. Incluimos también las especies relacionadas con la magia o los ritos o las que se considerase que asustaban a la gente o le afectasen de alguna otra manera, así como las plantas cuyas aplicaciones sugerían alguna actividad química no conocida. Además de las plantas que tenían efecto sobre el hombre, también tomamos nota de aquellas que, de alguna manera, se consideraban activas sobre los animales, incluyendo peces e insectos; las plantas con nombres locales que sugerían actividad biológica ("fruta del diablo" o "café de montaña", por ejemplo), y las plantas que se habían recolectado para análisis químicos.

Sin embargo, a medida que íbamos avanzando, fuimos eliminando todas las especies cuyos usos eran ya bien conocidos, quizás una quinta parte de las notas de interés. El grupo mayor que excluimos fueron las plantas que habían sido catalogadas en el *Dictionary of Economic Plants*, de J. C. Th. Uphof, con el que contrastábamos cada una de las anotaciones efectuadas. También eliminamos las llamadas "holotipos", que por definición son aquellas a las que se hace referencia en las publicaciones profesionales cuando se describe una nueva especie, así como cualquier especie cuyo nombre en latín ya pusiese de relieve propiedades medicinales o nutritivas. Y omitimos también las informaciones duplicadas sobre una misma especie de la misma zona geográfica.

Catalogamos los datos en el orden en que se encontraban las especies en las colecciones, familia por familia, con lo cual había una secuencia más o menos evolutiva. Tal ordenamiento servía para llamar la atención sobre características similares en especies próximas y sobre géneros y familias que tuviesen propiedades biodinámicas en común, todos los cuales podrían merecer atención especial. Las anotaciones comienzan con el nombre latino del ejemplar, citan el país en el que fue recolectado, el nombre del recolector y el número de la planta y luego transcriben las notas etnobotánicas u otras de interés.

Nos llevó cuatro años y medio revisar los dos millones y medio de ejemplares de plantas superiores montadas en pliegos de 30,5 × 43 cm. Terminamos con casi 6000 notas de interés referentes a más de 5000 especies y casi 2000 géneros; de unas 300 familias existentes en el herbario, más de la mitad estaban representadas en las notas. Las familias que estaban más ampliamente representadas eran las compuestas, que suponían el 15 por ciento de nuestras especies; las euforbias (8 por ciento); las legumbres (7 por ciento) y las mentas (6 por ciento). Desearía indicar que, en un trabajo similar hecho en otros herbarios, podrían estar muy representadas otras familias diferentes, ya que las colecciones tienden a estar especializadas en ciertas áreas; el 'Gray Herbarium' y el "Arnold Arboretum" son particularmente ricos en material de Asia Oriental, Norteamérica, Sudamérica y Sur del Pacífico.

Habíamos acometido el proyecto como una búsqueda de plantas medicinales, pero pronto se demostró que no se podía distinguir siempre entre drogas y alimentos; un té, por ejemplo, puede ser simplemente una bebida o puede ser un producto terapéutico. Además, para no desaprovechar ninguna sustancia biológicamente activa, habíamos decidido incluir todas las notas que indicasen que la especie era comestible o sugiriesen que era ingerida de alguna manera. En todo caso, no había razón para establecer una separación entre medicina y nutrición; la nutrición afecta a la salud, y nuevas plantas alimenticias podrían ser de gran valor potencial. Como resultado de lo que comenzó como una investigación de drogas vegetales, se produjo el hallazgo de posibles plantas alimenticias.

Ya que las anotaciones habían sido hechas por diferentes recolectores con muy diferentes intenciones e intereses, no había uniformidad en la clase de información etnobotánica en ellas incluida: iba desde indicaciones de que la planta era comestible (1255) a un informe sobre un tratamiento de la fiebre tifoidea. Por de pronto, catalogamos 922 plantas aromáticas, 407 llamadas medicinales, 284 venenosas, 241 recomendadas contra los trastornos intestinales, 190 consideradas con propiedades analgésicas y 116 que eran adecuadas para el tratamiento de lesiones. Otras etiquetas hacían referencia a plantas útiles para bebidas, tes, tónicos, astringentes y antídotos, o que se administran para el tratamiento de enfermedades o situaciones específicas.

Las plantas alimenticias entran dentro de varias categorías. Algunas especies pertenecen a géneros que no se sabía que incluyesen plantas de algún valor económico. Por ejemplo, se dice que *Burckella cocco*, de las Islas Salomón, produce "un fruto delicioso con una carne blanca, similar a una manzana, pero con un sabor muy distinto al de ésta". Encontramos muchos casos de especies poco conocidas, que pertenecen a géneros que tienen otras especies que sabemos son comestibles. Por ejemplo, un nuevo miembro, procedente del Perú, del género *Rubus*, que incluye las frambuesas y las zarzamoras. El género *Salvia* incluye muchas plantas de interés para las bebidas; a *Salvia hispanica* no se la conoce como una de ellas, pero encontramos muchas notas de América Central y de las Indias Occidentales que se refieren a su comestibilidad; un recolector dice que "se pueden producir fácilmente cosechas muy grandes". *Cyphomandra* incluye dos especies sudamericanas con frutos comestibles semejantes al tomate; nuestras notas incluyen una tercera especie, *C. splendens*, de la que se dice que se vendía como alimento en los mercados de Cuzco, en Perú. A veces una mera semejanza con un alimento puede ser una guía, como en el caso de *Uvaria confertiflora*, de Borneo, con un "fruto semejante a las bananas".

Algunas veces hay indicaciones de que una planta puede contener microelementos o vitaminas, enzimas o agentes antibióticos. Por ejemplo, "los nativos tienen la superstición de que si se dan fragmentos de este árbol a los cerdos, mezclados con el alimento, engordan rápidamente". Hay también plantas alimenticias "de emergencia" para el hombre y otros animales; hierbas de pasto preferidas y otras plantas que se dice son consumidas ávidamente por los animales; plantas que no se consumen directamente, pero que desempeñan un papel en la preparación de alimentos y plantas consumidas en cantidades extraordinariamente grandes. Otras especies de interés son aquellas consideradas de buen gusto, fragantes o aromáticas, aunque no se diga que sean comestibles; aquellas que son masticadas pero no ingeridas, y las asociadas con ritos agrícolas. Tales especies podrían cultivarse como complementos alimentarios.

Dentro del vasto campo de las posibles drogas vegetales, citaré algunos ejemplos de una página de nuestro trabajo tomada al azar. *Cologania angustifolia* var. *stricta*, de México, tiene "raíces que se consideran purgantes". *Erythrina rubinervia*, de El Salvador, tiene flores que "causan sueño si se comen en cantidad". En México, *E. lanata* "se usa como fuente de veneno entre los indios". El zumo de *E. variegata* var. *orientalis* "se usa para hacer un remedio contra la tos", en las Islas Salomón, y la cáscara de *E. subumbrans* se cuece en Filipinas "y se bebe contra la esplenomegalia" (dilatación del bazo). Las hojas de *Strongylodon lucidus*, de las Islas Salomón, "se calientan y luego se restriegan sobre los furúnculos". *Mucuna monosperma*, de Bombay, se toma "como medicamento, junto con otros ingredientes, para aliviar los espasmos agudos". Y *Calopogonium mucunoides*, de las Islas Carolinas, es un "remedio contra la debilidad general: ... se machacan las hojas y brotes jóvenes ... se exprimen en un cuenco ... se adiciona agua de coco y se bebe la mezcla".

Como ulteriores ejemplos, consideremos los agentes anticonceptivos potenciales y otras plantas asociadas con la reproducción. Encontramos una nota en relación a *Vochysia lomatophylla*, de Perú, que es "quizás usada por la tribu Campa como anticonceptivo". Hay otras plantas que se administran para la regulación de la menstruación, pero que pueden actuar también como reguladores de la fertilidad. Algunas especies se citan como ayuda en el parto; así: *Ocimum sanctum*, de Filipinas. Uno se pregunta si esta planta de la familia de las mentas puede tener o no un constituyente psicofarmacológico, como lo tienen, según datos recientes, los géneros próximos *Nepeta* (hierba gatera) y *Salvia*. *Ocimum* podría desarrollar un efecto similar a la escopolamina, que es un alcaloide hipnótico de origen vegetal.

Las muestras de herbario se conservan largo tiempo, si están bien cuidadas. (Se cuenta que cuando la colección Linneana estaba siendo fotografiada en 1939, un fotógrafo se hirió en un brazo con una ortiga que había sido montada casi 200 años antes; le produjo la misma clase de ampolla que produciría la planta viva.) Por tanto, como complemento a la lectura de las etiquetas, se pueden hacer "ensayos rápidos *in situ*" en busca de determinados constituyentes químicos en pequeños fragmentos de la planta, cuidadosamente seleccionados, con el fin de conservar la integridad del ejemplar. Por este método se pueden examinar suficientes ejemplares representativos de un herbario extenso como para determinar la posible presencia de un compuesto químico en grupos determinados de plantas. Los alcaloides estables pueden dar prueba positiva después de 125 años en un herbario. Hace poco se obtuvo resultado positivo en una prueba de alcaloide hecha sobre una hoja de coca, sepultada con una momia peruana alrededor de 500 años d. C.

Además de la nutrición y la medicina, al menos otras dos disciplinas pueden beneficiarse de las investigaciones de herbario: la taxonomía vegetal y la entomología. Uno de los enfoques más recientes de la taxonomía es la quimiotaxonomía, que persigue la determinación de las relaciones entre las plantas por medio de su química comparada. La información sobre medicina popular obtenida de las notas de campo, unidas a los ejemplares de herbario, refleja generalmente la química de la planta y puede servir para seleccionar las especies en las cuales pueden encontrarse probablemente determinados compuestos químicos. Los estudiosos de la etnotaxonomía, que investigan los sistemas por los cuales los pueblos primitivos clasifican sus plantas, pueden servirse también de nuestras notas, que incluyen un gran número de nombres indígenas.

Las actuales investigaciones en entomología han hecho grandes progresos al esclarecer las interrelaciones químicas entre plantas e insectos. Parece que ciertos insectos toman productos químicos de las plantas que comen, acumulan dichos compuestos y posteriormente los emiten para defensa, como "agregado de feromonas", o quizá como reclamos sexuales. Ciertas orquídeas tropicales americanas producen unos compuestos de olor suave y delicioso, biológicamente activo, que atraen a las abejas macho, los cuales no los comen pero los recogen, los almacenan probablemente pueden metabolizarlos. Estas abejas viven más que otras abejas macho, y existen ciertos que otras macho, y existen ciertos indicios de que mueren si se les priva de tales compuestos. ¿Podría haber aquí alguna pista sobre los procesos de envejecimiento en general? Las notas de herbario que registran inusitadas relaciones entre insectos y plantas podrían resultar valiosas en este vasto campo de investigación.

Deberían llevarse a cabo exploraciones semejantes a las nuestras en todos los herbarios del mundo. Algo parecido ha realizado después Teresa Fossi Innamorati, en el Herbario Tropical de Florencia; mis colegas y yo estamos ocupados ahora en otra investigación en el New York Botanical Garden. La mayoría de las especies, incluyendo aquéllas que las sociedades primitivas han considerado útiles desde siempre, nunca han sido examinadas desde este punto de vista y con las técnicas de la ciencia moderna. Existe una razón para darse prisa. Muchas especies están en peligro de extinción, como lo están los pueblos y las costumbres que albergan información sobre ellas, conforme se van expandiendo las sociedades avanzadas.

Bioquímica de la subnormalidad

Muchos casos de subnormalidad pueden evitarse. Urge aplicar los recursos científicos disponibles, desde medidas de índole sociológica hasta una mayor atención obstétrica y diagnóstica, y favorecer el desarrollo de la investigación

Federico Mayor

Dice el viejo refrán castellano: "más vale prevenir que curar". Para prevenir, es necesario conocer. El mejor índice de que algo se domina es que pueda evitarse. Y ello es así en todos los órdenes de la vida, pero especialmente en aquel que incide de una forma directa en el bienestar del hombre: su salud. Desde su misma aparición sobre la tierra, el hombre ha puesto toda su imaginación para mejorar los recursos con que alimentarse y evitar o paliar las enfermedades que le afligen. Estos recursos se han ido incrementando progresivamente. Sin embargo, la contemplación panorámica del mundo de hoy pone de manifiesto una grave desproporción entre el progreso técnico alcanzado para satisfacer aspectos secundarios de la vida humana y la adecuada –y general– aplicación de los conocimientos disponibles para la atención nutritiva y sanitaria exigibles, especialmente cuando, por no actuar a tiempo y concentrar en estas grandes prioridades todos los esfuerzos al alcance, se producen lesiones e incapacidades permanentes.

Introducción

Los nuevos medios de diagnóstico y tratamiento, tanto medicamentoso como quirúrgico, han permitido un descenso espectacular de la morbilidad y mortalidad perinatal y han incrementado la incidencia de casos en los que la raíz genética es fundamental. Por selección natural, muchos de estos nuevos seres precariamente dotados para hacer frente a su ambiente, a sus circunstancias, desaparecerían. Hay deficiencias genéticas que se han compensado como "acto normal" de la propia trayectoria de la especie humana: el hombre es incapaz de sintetizar algunos factores vitamínicos indispensables, por carecer –a veces muy específicamente– de los enzimas (o de alguno o algunos de ellos) necesarios para su biosíntesis. Pero estos factores se hallan en el ambiente normal del hombre –en caso contrario no podría sobrevivir–, y éste consume los alimentos que los contienen. Lo mismo pasa con los antígenos "naturales". Sin embargo, el desarrollo tecnológico ha producido una nueva serie de antígenos que no se hallaban en el entorno, es decir, han aparecido unos nuevos "agentes de selección" artificiales o no naturales que, al actuar sobre los genotipos, ponen de manifiesto características no vislumbradas anteriormente.

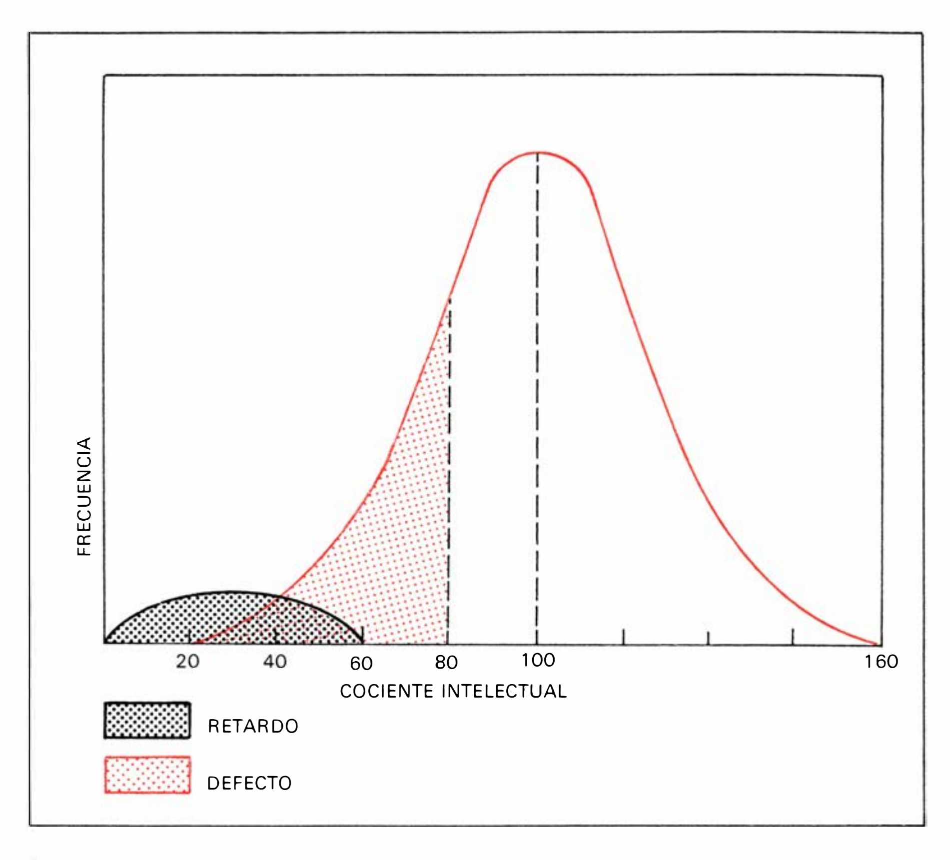
Todas las alteraciones genéticas, hereditarias por tanto, fueron en su momento adquiridas. Algún antecesor sufrió una mutación que lesionaba en mayor o menor grado la funcionalidad de un sistema enzimático, lo cual determinaba una inhabilidad parcial, compatible con la vida.

Los agentes mutágenos, al actuar aislada o conjuntamente, modifican el mensaje genético, en el caso de las enfermedades hereditarias, o interfieren con mayor o menor especificidad algunas etapas de la morfogénesis (que es el desarrollo de la forma y la estructura del organismo) y de la diferenciación (fase durante la cual las células no especializadas se van especializando en sus funciones características), en el caso de las enfermedades no hereditarias. Se diferencia de este modo lo que podríamos denominar "causas primeras" o iniciales, y las que conducen a malformación en el desarrollo de un ser determinado debido a la acción de un agente determinado y en un determinado momento de su desarrollo. En resumen, las reacciones iniciales de un sistema en desarrollo frente a cualquier interferencia se traducen en acontecimientos morfogenéticos o bioquímicos que conducen, por último, a anormalidades estructurales o funcionales. A nadie escapa la íntima relación que existe, a su vez, entre estructura y función y acontecimientos bioquímicos y morfogenéticos. Cualquier desviación del proceso normal de diferenciación conduce a manifestaciones funcionales o estructurales, y es en la descripción de dichas manifestaciones más que en su origen en donde pueden establecerse claras separaciones, útiles a los efectos de estudio e investigación así como para su adecuada consideración clínica y tratamiento.

Los términos "teratológico" (monstruoso) y "malformación" se aplican a anomalías estructurales. Los términos "alteraciones congénitas, alteraciones metabólicas hereditarias", etc. suelen adscribirse a los desórdenes funcionales. Los defectos teratológicos se han producido antes del nacimiento, aunque la descripción de tales anormalidades, si el feto es viable, es decir, si tiene capacidad para subsistir, corresponde al aspecto que tengan los nuevos seres en el momento de nacer. Todo desarrollo anormal tiene su causa remota, seguramente, en algún factor ambiental nocivo. Los defectos congénitos pueden deberse a alteraciones en los genes de los predecesores, transmitidas a lo largo de sucesivas generaciones. Si el efecto ambiental tiene lugar después de la fecundación, las anomalías que induce afectan exclusivamente a este ser, en el supuesto de que el feto sea viable. Las que modifica-

ron la calidad genética, en cambio, son transmisibles, ya que el nuevo ser hereda una carga genética con características alteradas. Los conocimientos actuales sobre este particular son todavía incipientes, pero se sabe que las radiaciones, por ejemplo, pueden inducir tanto alteraciones hereditarias como no hereditarias. Algunos productos químicos pueden, por otra parte, causar efectos muy distintos según la especie afectada, el tiempo de gestación, etc. También algunas infecciones maternas padecidas durante la gestación, especialmente durante el periodo embriogénico, producen alteraciones teratológicas graves, como en el caso de la rubéola, mientras que otras tienen escaso o nulo efecto.

En última instancia, no existen diferencias entre las malformaciones de sintomatología muy aparente y aquellas otras poco aparentes. Todo afecta a la forma y sólo depende del nivel de observación en el que nos situemos a la hora de reconocer la anomalía. La morfología depende de la unidad de medición: en último término, es la estructura molecular, la anatomía de las moléculas, la que se halla alterada.



SUBNORMALES PROFUNDOS son los individuos cuyo coeficiente intelectual (C. I.) es inferior a 25; los "subnormales severos" se hallan entre 25 y 39; "moderados" entre 40 y 54; "ligeros" entre 55 y 70; "límites" o "border line" entre 70 y 80. El déficit intelectual que carece de base orgánica patente se denomina retardo mental; por el contrario, el defecto mental va acompañado de una clara sintomatología, con evidentes alteraciones orgánicas más o menos graves.

Personalidad bioquímica

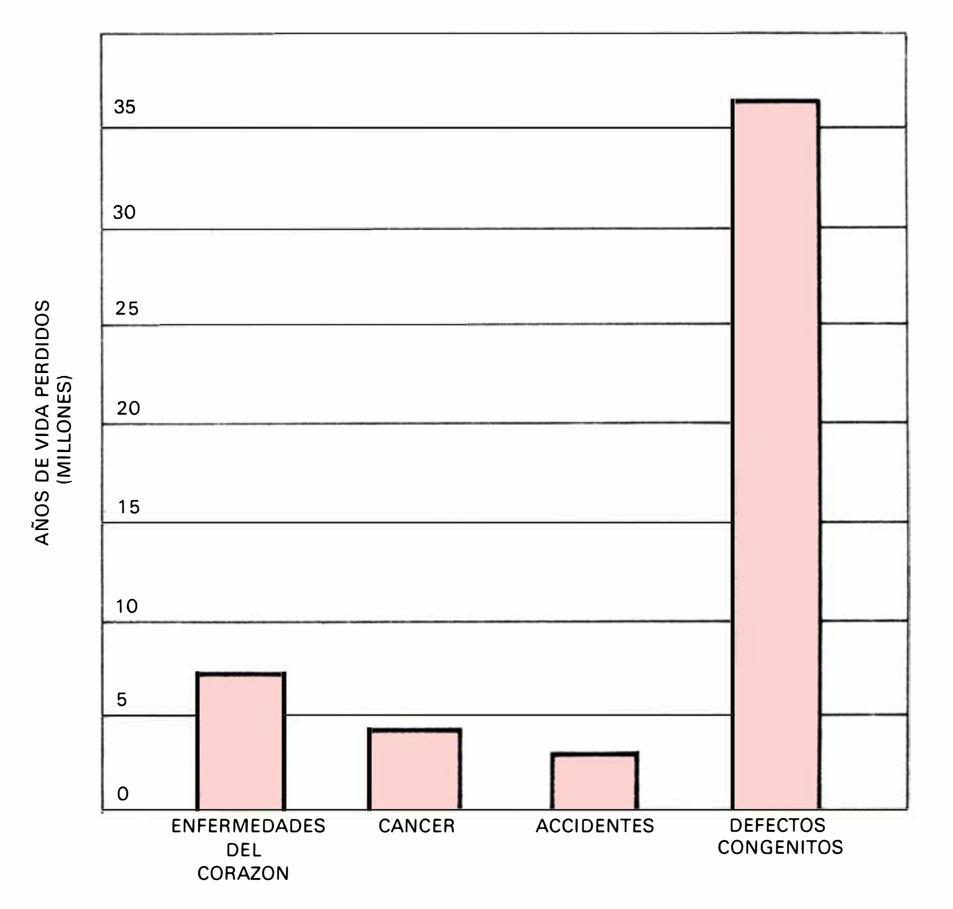
El mayor legado de la evolución, ha escrito Scriver, es la fantástica diversidad genética del hombre, que pone de manifiesto las bases biológicas de la individualidad de los seres humanos. Permite distinguirnos a nivel molecular igual que podemos distinguirnos –ya que no constituye sino la exteriorización de nuestra íntima realidad molecular– por nuestro aspecto, por nuestra altura, por nuestras caras. El polimorfismo bioquímico, lectura del genético, ha sido decisivo en el planteamiento de los nuevos criterios de individualidad química. Individualidad tan certeramente presentida por Garrod cuando puso de manifiesto la primera metabolopatía congénita: la alcaptonuria, enfermedad hereditaria debida a la ausencia de un gen que regula la oxidación del ácido homogentísico.

Frente a esta gran diversidad, la bioquímica nos ofrece la extraordinaria unidad de procesos básicos en todos los seres vivos. Así, todos ellos proceden con la misma moneda energética, tienen los mismos esquemas de degradación de los combustibles orgánicos, etcétera.

La individualidad se pone de manifiesto como una forma de incompatibilidad con el entorno, en mayor o menor grado. Esta es la razón de las reacciones alérgicas. Es evidente que la solución puede ser: *a)* desplazar del ambiente al componente frente al cual se está inerme genéticamente; *b)* desplazarse uno de este ambiente; *c)* dotar al individuo de la capacidad de que carece. El gradiente de gravedad será proporcional, de un lado, a la normalidad de la presencia en el "ambiente" del componente que descubre la individualidad y, de otro, de la importancia bioquímica de los procesos alterados. De hecho, en el esquema de Scriver no existe otra diferencia entre la galactosemia (que pone como ejemplo de enfermedad de origen genético) y la deficiencia en el enzima glucosa-6-fosfato deshidrogenasa (que figura como ejemplo de origen multifactorial) que la "normalidad ambiental" del componente que pone de relieve estas peculiaridades bioquímicas: la galactosa forma parte de la lactosa, azúcar de la leche y, por tanto, es un elemento esencial de la leche en el recién nacido; en cambio, la deficiencia en glucosa-6-fosfato deshidrogenasa se revela habitualmente sólo cuando los seres que son deficientes en este enzima, ligado al cromosoma sexual X, ingieren determinados alimentos –como las "habas de mayo"– o determinados medicamentos de naturaleza oxidante. Como vemos, todo depende de que el agente "provocador" sea más o menos normal, de que se presente en una etapa más inicial o tardía del desarrollo.

Viabilidad. Vulnerabilidad

La íntima relación entre las características de la dotación genética y el ambiente nos lleva a considerar, siempre al hilo de la personalidad bioquímica, los conceptos de viabilidad, vulnerabilidad y, como conjunción de ambos, los intervalos que comprende la salud (es decir, "lo normal") y la enfermedad (lo patológico). Sólo lo que es viable es objeto normalmente de estudio y consideración; sólo vemos aquellas alteraciones que evidencian una deficiencia genética cuando el ambiente las delata. La posibilidad de manifestación será tanto mayor cuanto más cerca se halle del "umbral de alarma", del límite de tolerancia. Este es el caso de muchos transportadores de proteinopatías heterocigotos (dotados de dos alelos distintos para cada gen), que sólo exhiben sus limitadas capacidades funcionales cuando les sometemos a pruebas de "sobrecarga"



LA DIFERENCIACION ANORMAL, dada su alta incidencia, ha figurado ya en los últimos años entre las principales enfermedades que afligen a la humanidad. Si la comparación se establece sobre la base de años de vida, tal como se representa aquí, la importancia relativa de las alteraciones congénitas o adquiridas presentes al momento de nacer alcanza un especial relieve.

en el componente para cuya confrontación se hallan solo *parcialmente* capacitados. La respuesta es parecida, pero no idéntica, según sea mayor o menor la eficacia de la proteína enzimática para el cumplimiento de su función. Y es que hay un intervalo de normalidad como lo hay de patogeneidad. Las mutaciones que sufre la secuencia genética y, por tanto, la estructura proteica pueden: *a)* conducir a una estructura totalmente inhábil para el cumplimiento de la actividad que les es propia; *b)* no afectar la actividad de la proteína, o hacerlo en un grado que resulta imperceptible, al menos en condiciones normales, al no alterarse substancialmente su disposición espacial; *c)* alterar de tal modo la secuencia monopeptídica (secuencia de aminoácidos unidos por enlace peptídico) que se produzca una deficiencia de mayor o menor repercusión, dentro de un amplio espectro de gravedad. En muchos casos, la alternativa *c)* resulta muy similar a la de los heterocigotos o transportadores, y con frecuencia no se pone en evidencia sino cuando incide extemporáneamente un entorno especialmente "exigente" en relación a la actividad deficitaria.

Las lesiones cerebrales que se producen en los primeros meses de vida se explican por el alto grado de vulnerabilidad que tiene el cerebro durante este periodo. Las mismas circunstancias no producirían, más adelante, el daño permanente, las secuelas irreversibles que producen en esta etapa de extraordinaria inmadurez y fragilidad del cerebro. También otras células (del hígado, riñón, etc.) pueden afectarse en una situación caracterizada por el acúmulo de sustancias que normalmente se hallan en menores proporciones y que resultan tóxicas para el organismo. Imaginemos una hiperbilirrubinemia en el neonato. El hepatocito (célula del hígado) queda afectado. Las neuronas también. Pero el hepatocito se renueva y las neuronas no. Por esta razón, el daño producido directamente sobre las células cerebrales (neuronas y glías) es irreparable. Por constituir una "dotación de por vida", el sistema nervioso se halla fuertemente protegido por unas capas "aislantes" de mielina. Durante este tiempo, en consecuencia, una interferencia en la formación de mielina puede constituir una causa de lesión cerebral. También pueden resultar lesionadas las células todavía "desnudas".

El momento de mayor vulnerabilidad es, desde luego, aquel en que las neuronas se hallan más expuestas a la agresión de agentes patógenos, lo que no excluye que, una vez ya protegidas por las capas de mielina y alcanzada la plena funcionalidad de la barrera hematoencefálica, pueda producirse una lesión permanente por alteraciones directas y graves de las estructuras del sistema nervioso central, tales como infección de las meninges, heridas traumáticas y otras.

Como vemos, en la consideración de la fisiopatología del desarrollo cerebral no deben considerarse por separado los distintos componentes celulares que integran el encéfalo. Bien al contrario, el binomio glía-neurona debe ser considerado en su conjunto, pues es conjuntamente como realizan, por ejemplo, todo el proceso de protección axonal: las glía rodean los axones y son las responsables, en último término, de la mielinización de los mismos.

En la rata, la mielinización del cerebro tiene lugar entre 10 y 21 días *después* del nacimiento. El término del proceso coincide con el destete. En el hombre, en cambio, la mielinización se inicia a los cuatro meses de gestación, y no se completa sino *hasta* los dos años de vida. La mielogénesis no se produce de una manera uniforme en el sistema nervioso, ni su velocidad es la misma en las distintas partes del cerebro; debe tenerse muy en cuenta al investigar los efectos de sustancias interferentes ya que, según el momento que se considere, unas áreas estarán prácticamente mielinizadas mientras que otras se hallarán iniciando el proceso de protección. En la consideración de la patogénesis, hay que considerar, asimismo, el efecto directo que sobre las células no protegidas puedan tener los agentes que alteran el desarrollo normal del sistema nervioso. En otras palabras, dichos agentes pueden impedir que se forme la protección de la célula nerviosa, pero también pueden actuar directamente sobre ésta en un momento en que se halla expuesta a la acción de dichos agentes, situación que no se presenta después, una vez consolidado todo el sistema "aislante" de naturaleza mielínica.

El proceso de desarrollo prenatal debe considerarse en la totalidad del sistema implicado, es decir, el conjunto madre-feto, de la misma manera que el desa-

rrollo neonatal y posnatal debe juzgarse considerando como sistema el conjunto ser vivo-entorno. A su vez, tanto en un caso como en otro, hay que tener muy presente la posible interferencia multifactorial y las interacciones correspondientes entre las consecuencias que de ella se derivan. Al considerar en conjunto los sistemas mencionados pueden establecerse con la precisión justa los diversos grados de sensibilidad –que dependen de la protección efectiva que otorga el entorno (bien sea la madre o el ambiente) al ser cuyo desarrollo se considera– o, lo que es lo mismo, el distinto grado de vulnerabilidad del ser en desarrollo, en sus diferentes etapas. La mayor vulnerabilidad, en lo que hace referencia a desviaciones estructurales, tiene lugar durante el periodo de la organogénesis, es decir, desde los días 18-20 hasta los días 55-60. Una vez completada la organogénesis, la susceptibilidad a los defectos de índole anatómica disminuye considerablemente, aunque pueden producirse otras desviaciones estructurales hasta que se completa la histogénesis, al final del periodo fetal. Las desviaciones durante este periodo son principalmente de índole funcional. Por último, las interferencias que conducen al retraso mental tienen una mayor incidencia –excepto en los casos en que las actividades enzimáticas deficitarias son requeridas para el desarrollo intrauterino o en aquellos otros en que la malformación, de origen genético o no, incide precozmente en la morfogénesis– en el periodo perinatal y neonatal, debido al singular desarrollo del cerebro, especialmente en lo que se refiere a la producción de las estructuras que protegen las neuronas. El mayor grado de vulnerabilidad coincide, evidentemente, con la mayor letalidad embrionaria. Piénsese, en todas estas reflexiones, que estamos tratando del momento en que se *visualiza* una acción determinada sobre el proceso de diferenciación y desarrollo. De tal manera que puede ocurrir –y de hecho ocurre– que las malformaciones reconocibles en un momento dado del proceso tengan su origen en la misma dotación genética o en una interferencia muy inicial de su normal expresión, que no se pone de manifiesto hasta más adelante. Precisamente, cuando es necesaria aquella actividad precozmente inhibida o alterada. Esta visualización o manifestación puede tener lugar tanto durante el desarrollo intrauterino como en el periodo neonatal, cuando al adquirir la vida autónoma cesa la acción protectora de la madre.

La prevención primaria es la que evita, en lo posible, la aparición de subnormalidad. La que alivia o cura los transtornos ya presentados se denomina prevención secundaria. La reducción del grado de incapacidad se incluye en la denominación de prevención terciaria.

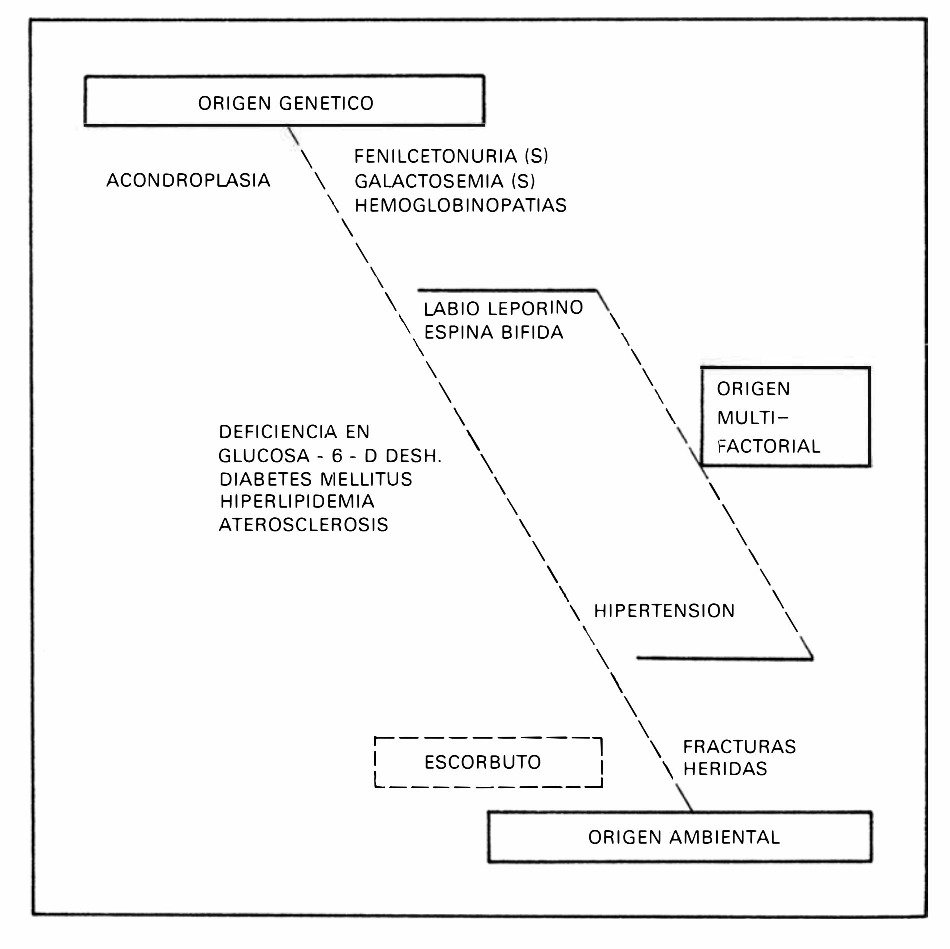
Alcance y frecuencia de la subnormalidad

Muchos autores distinguen entre retardo mental (déficit intelectual sin base orgánica clara) y defecto mental (déficit intelectual con base orgánica probada). Así como en el defecto mental aparecen alteraciones orgánicas graves, en el retardo mental es importante considerar la etiología socio-cultural y orgánica poligénica.

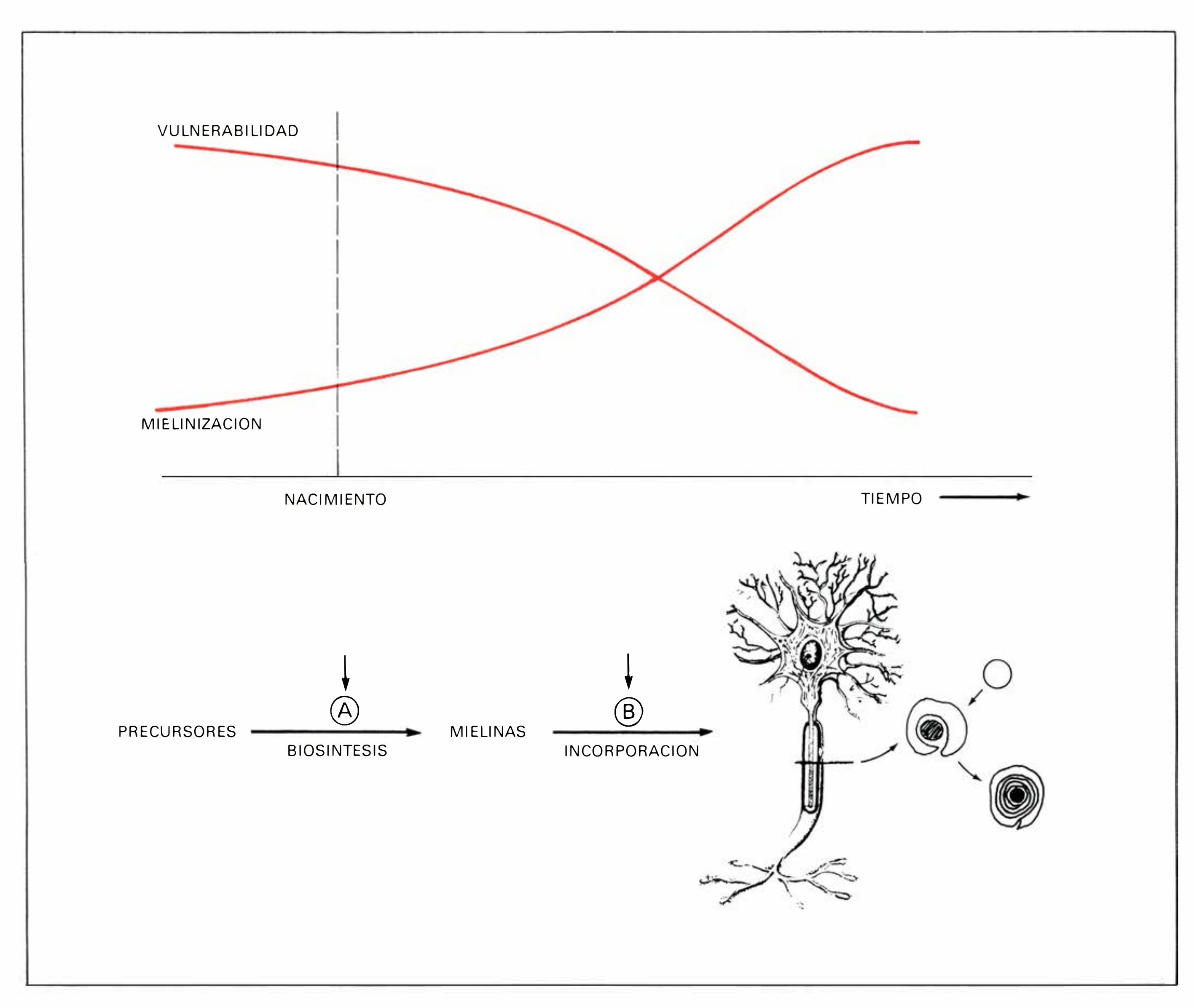
Quienes presentan un cociente intelectual mayor de 50 por ciento, son educables. Cuando el cociente intelectual se halla comprendido entre 25 y 50 por ciento, son entrenables. Si es inferior a 25 por ciento, son asistibles. Interesa muchísimo la medición periódica del cociente intelectual, a fin de que, al personalizar la atención educativa y formativa se facilite en cada momento la atención sanitaria y cultural correspondiente.

El retraso mental es un fenómeno biológico, pero también socio-cultural. No podemos considerar la producción de la lesión fuera del ambiente que la provoca o manifiesta. Lo biológico y lo social aparecen como realidades conjuntas, interactuantes e inseparables. Este tema requiere, en consecuencia, una consideración global.

Hasta hace pocas décadas –un tiempo insignificante si se compara con el proceso evolutivo– se adquiría la autonomía biológica en un mundo "natural", en un contexto coherente con las dotaciones del nuevo ser. Ahora no. Ahora se nace a un mundo progresivamente artificial y, por tanto, con unas nuevas exigencias, que deben satisfacer ineludiblemente. Las radiaciones, los



ORIGEN DE LAS ALTERACIONES a partir del esquema de Scriver (1976). Existe una íntima conexión entre los orígenes representados: enfermedades monogénicas y cromosómicas *(arriba)*, enfermedades multifactoriales por concurrencia de factores genéticos y ambientales *(centro)* y alteraciones de origen ambiental *(abajo)*. Toda alteración de origen genético fue adquirida, en un momento dado, por algún antecesor a través de una mutación. La mutagénesis es favorecida con frecuencia por las características del entorno (radiaciones, productos químicos, etcétera). La individualidad bioquímica se pone de manifiesto por la calidad del ambiente, según sea el grado de compatibilidad. Las propiedades que caracterizan a cada ser están sujetas a múltiples cambios. El gradiente de gravedad de cada alteración resultará ser, en consecuencia, extraordinariamente variado.



LA VULNERABILIDAD CEREBRAL decrece a medida que aumenta el grado de mielinización. En la rata, la biosíntesis de mielinas es un proceso posnatal. En el hombre se inicia alrededor de los cuatro meses de gestación, pero no culmina hasta varios meses después del nacimiento. En la cobaya, en cambio, la mielinización es prácticamente total al nacer. Dado que las neuronas no se renuevan, las lesiones producidas durante su diferenciación y protección son permanentes. El proceso de mielinización puede interferirse en varios niveles: *A*, en la formación de las mielinas a partir de sus precursores; *B*, en su incorporación alrededor del axón. Mientras el sistema nervioso se está recubriendo de las capas aislantes, pueden resultar lesionadas las células todavía "desnudas". Ello explica –junto a la existencia de procesos bioquímicos específicos del neonato y a una menor funcionalidad de la barrera hematoencefálica, que permite el acceso al cerebro de sustancias que más adelante serán cuidadosamente seleccionadas– que durante los primeros meses de vida el cerebro humano resulte irreversiblemente lesionado por circunstancias metabólicas que no le afectan ulteriormente, una vez consolidadas las estructuras que lo integran. Las células de Schwann rodean el axón.

productos tóxicos, el grado creciente de contaminación... son índices del deterioro "no natural" de la biósfera que, si afectan la salud de los hombres en general, afectan en mayor medida, lógicamente, a los más débiles, a los menos protegidos frente a estos factores, a los que se hallan en los estadios iniciales de su propia formación, a los que se hallan en el trance crítico de su advenimiento. Hoy nacen muchos niños que no hubieran nacido sin la ayuda "no natural" facilitada a la madre durante la gestación. Por ejemplo, de cada 6 fetos afectos de cromosomopatías (alteraciones en los cromosomas) sólo uno llega a nacer en la actualidad, y puede asegurarse que estos fallos de la "aduana del control de calidad" se incrementan por los tratamientos que se aplican a la madre.

Por lo menos el 4 por ciento de los nacidos vivos sufren una anomalía total o parcialmente genética (el 1 por ciento son de índole cromosómica). En los Estados Unidos, la incidencia del retraso mental es del 3 por ciento de los nacidos vivos. En España se calcula que existen 350 mil subnormales, sobre 35 millones de habitantes, lo que equivale al 1 por ciento de la población total. Claro está que se trata de cifras muy aproximadas, ya que lo primero que debe homologarse es el concepto de "subnormal" y, por tanto, referirlo a la población que realmente comprende.

En niños de menos de 5 años, la distribución del grado de subnormalidad mental puede establecerse, de acuerdo con una frecuencia de subnormalidad del 2,8 por ciento y según datos recogidos por Sánchez-Villares, como sigue: ligeros 845, moderados 112, severos 89 y profundos 82.

Comparemos estos datos con las cifras correspondientes a minusvalías de distinta naturaleza, que expresan el panorama de un país avanzado, Francia, en una fecha relativamente reciente, 1968.

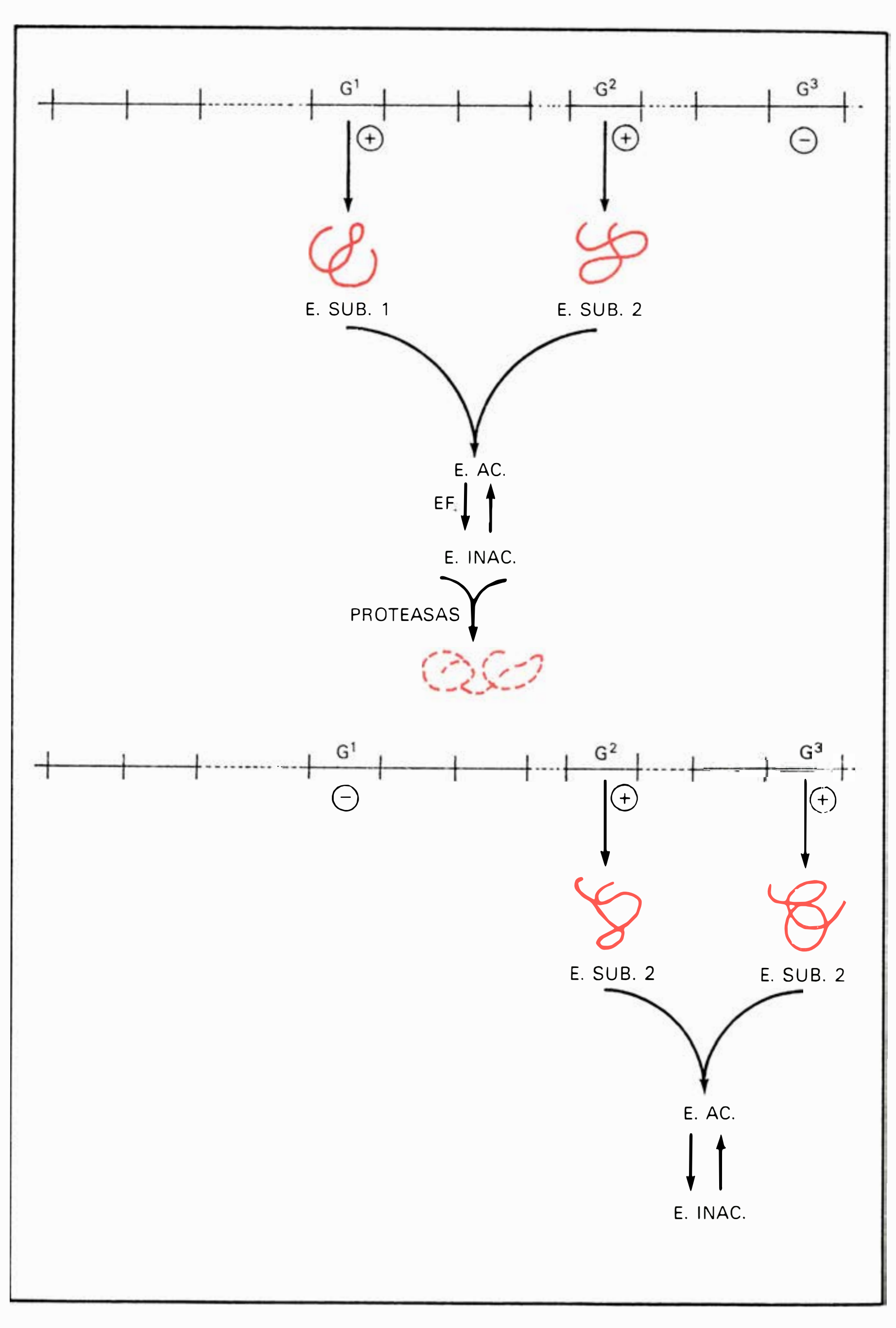
El total de nacidos vivos en ese año ascendió a 833.000; de ellos, 26.300 padecían subnormalidad mental, lo que suponía un 3,16 por ciento sobre nacidos vivos; 2.900 padecían otras anomalías del sistema nervioso (parálisis cerebral, etcétera), lo que implicaba el 0,35 por ciento; 7.200 sufrían defectos congénitos del corazón, pulmones, tracto digestivo y sistema urogenital, lo que equivalía al 0,86 por ciento y 3.600, o sea el 0,43 por ciento, padecían deformidades físicas, bucales y genitales. Había, pues, 40.000 niños minusválidos (que sobrevivieron no menos de un año), es decir, el 4,80 por ciento del total de nacidos vivos.

Hay que destacar, además, la gran importancia de las alteraciones presentes al nacer, si se consideran los años de vida perdidos en comparación con otras enfermedades de gran incidencia, tales como el cáncer, los accidentes o las cardiopatías. Podemos agrupar, en razón del periodo de ocurrencia, los factores que originan estas alteraciones: factores preconcepcionales (genéticos), factores prenatales (infecciones, agentes químicos, físicos, nutricionales, inmunológicos, etcétera), factores perinatales (asfixia, lesión en el parto, prematuridad, etcétera) y factores postnatales (infecciones, traumatismos, factores nutricionales, etcétera).

Predisposición

Como ha indicado Childs, si la selección genética masiva tuviera que limitarse a los ensayos actualmente empleados, seguiría constituyendo una parte insignificante de la atención sanitaria. Es cierto que la relación de metabolopatías congénitas, cuya incidencia suele ser baja, junto con algunas anormalidades cromosómicas, de mayor incidencia, seguirían aportando interés práctico y científico a la realización de estas determinaciones. Pero, sin duda, es en las enfermedades comunes y crónicas de origen genético todavía desconocido en donde hallamos el horizonte más prometedor por su importancia cuantitativa en la patología, para continuar profundizando en el conocimiento de la variabilidad genética y procurar la prevención de trastornos de tan alta incidencia. El descubrimiento de la magnitud del polimorfismo genético humano y la evidente relación de algunas enfermedades comunes con ciertos genes polimórficos confieren especial énfasis a la importancia de profundizar en el conocimiento de los genes responsables. Harris y Hopkinson han puesto de manifiesto la existencia de alelos polimórficos de no menos del 30 por ciento de los *loci* humanos, siendo probablemente mayor el porcentaje de genes que, si bien "inocuos" cuando se les considera aisladamente, pueden constituir el origen de múltiples enfermedades humanas si se combinan con otros genes o se expresan en circunstancias especiales.

Hoy se juzga que la predisposición es de origen multifactorial, si bien se acepta que es de naturaleza muy heterogénea, tanto fenotípica como genéticamente, con unas características que se está intentando clasificar en síndromes distintos, de tal manera que se los pueda investigar mediante escrutinio genético y molecular. La mayoría de los casos parecen ser el resultado de influencias multigénicas y ambientales, mientras que en ciertos individuos y familias algunas de estas alteraciones cumplen claramente la normativa mendeliana, unifactorial. La diabetes, la gota y la hipertensión son ejemplos de las predisposiciones que comentamos. ¿Cómo contribuyen estos ge-



INDUCCION DE UN ENZIMA ACTIVO integrado por dos subunidades. La cadena polipeptídica (subunidad 1) que procede de la lectura del gen estructural 1 (G^1) se une con la subunidad 2 (procedente de G^2) para formar el enzima (1+2) activo, poco regulable, y que, además, resulta ser necesario en las primeras etapas del desarrollo. Más adelante, se induce G^3, se reprime G^1, y se forma el enzima (2+3), altamente regulado, que es característico del metabolismo adulto.

nes relativamente frecuentes a la aparición de enfermedades tan comunes? Una posibilidad es la de los "transportadores" o heterocigotos, que contienen una sola "dosis" del gen correspondiente. En los homocigotos, caso de ser viables, la enfermedad es grave. En los heterocigotos la deficiencia puede pasar inadvertida o bien mostrarse en virtud de los cambios que se producen con la edad o en virtud de la influencia de factores externos cuya "concentración" ponga de manifiesto una singularidad que, en caso contrario, hubiera permanecido oculta. Así, se ha sugerido que el retraso mental, normalmente leve pero con una clara predisposición familiar, que se origina como consecuencia de una malnutrición materna moderada se debe a que el feto es heterocigótico para alguno de los genes relacionados con las alteraciones del metabolismo de los aminoácidos, unidades integrantes de las proteínas. Como vemos, los "transportadores" quedan al descubierto –al igual que hacemos en el laboratorio con las pruebas de sobrecarga– cuando las exigencias metabólicas superan las posibilidades de acción de una sola dosis genética.

Es ciertamente difícil realizar estos estudios sobre una base prospectiva, a saber, haciendo los análisis en el periodo neonatal o en la infancia y viendo si, más adelante, se produce alguna alteración que pueda atribuirse a "predisposición" genética. Pero podría hacerse perfectamente sobre una base retrospectiva, de manera que pudieran delimitarse grupos de alto riesgo, que se llegara a establecer una relación fundamentada entre una alteración patológica y una insuficiente dotación genética (traslucida por el acúmulo de algunas sustancias) y, lo que es más importante, facilitara la adopción de medidas que permitieran evitar o paliar la incidencia de algunas enfermedades "comunes". Así, las hiperlipemias, asociadas tan frecuentemente con los infartos de miocardio, pueden ser debidas a un condicionamiento genético, con independencia de otras causas (ingesta, por ejemplo) que puedan conducir, asimismo, a un incremento peligroso en la cantidad de colesterol y triglicéridos circulantes. Según las características de sus correspondientes dotaciones genéticas, algunas personas mostrarán una especial sensibilidad a dicho incremento mientras que otras serán capaces de afrontarlo perfectamente. Si se lograra realizar las determinaciones correspondientes en los familiares de los enfermos cardíacos, con especial énfasis en su descendencia, se podría obtener conclusiones de gran relieve sobre la variabilidad genética y la proclividad hereditaria hacia una alteración determinada y se sentarían unas bases firmes para la prevención de uno de los grandes orígenes de la morbilidad y mortalidad actuales.

El neonato. Primero, sobrevivir

El neonato se ha configurado, en buena medida gracias a los conocimientos derivados del polimorfismo isoenzimático (multiplicidad de formas de un mismo enzima), como una entidad biológica perfectamente deslindada de las que le anteceden y suceden, con unas dotaciones isoenzimáticas en perfecta armonía con las funciones de puesta en marcha y adecuación de actividades hasta este momento innecesarias –o necesarias de otro modo– y dirigidas todas ellas a la sobrevivencia. Este polimorfismo bioquímico del recién nacido permite, a través de las formas isoenzimáticas menos regulables, la movilización general de todos los resortes requeridos para la consolidación de la recién adquirida autonomía energética, metabólica, biológica en suma. A continuación, pasadas las primeras horas o los primeros días, con un ritmo perfectamente establecido en cada caso, de acuerdo con las conveniencias funcionales, disminuye progresivamente la lectura de los genomas que conducen a dichas formas isoenzimáticas y se induce la síntesis y activación de las formas más regulables, bioquímicamente más complejas y perfectas, que deben sustituirlas. Este mecanismo –sustitución, al ritmo debido, de unas formas isoenzimáticas por otras– caracteriza la ontogénesis normal. El desarrollo neoplásico aparece como una brusca alteración de la diferenciación celular. Y así, muchos tejidos cancerosos presentan un cuadro de isoenzimas similar o idéntico al de los tejidos embrionarios. Este comportamiento se debe, al parecer, a la represión o no activación en el momento oportuno de los genes que normalmente se expresan en la célula adulta diferenciada, al tiempo que no se reprimen (o son activados) los genes que dirigen las actividades de los primeros estadios de diferenciación celular.

Cromosomopatías

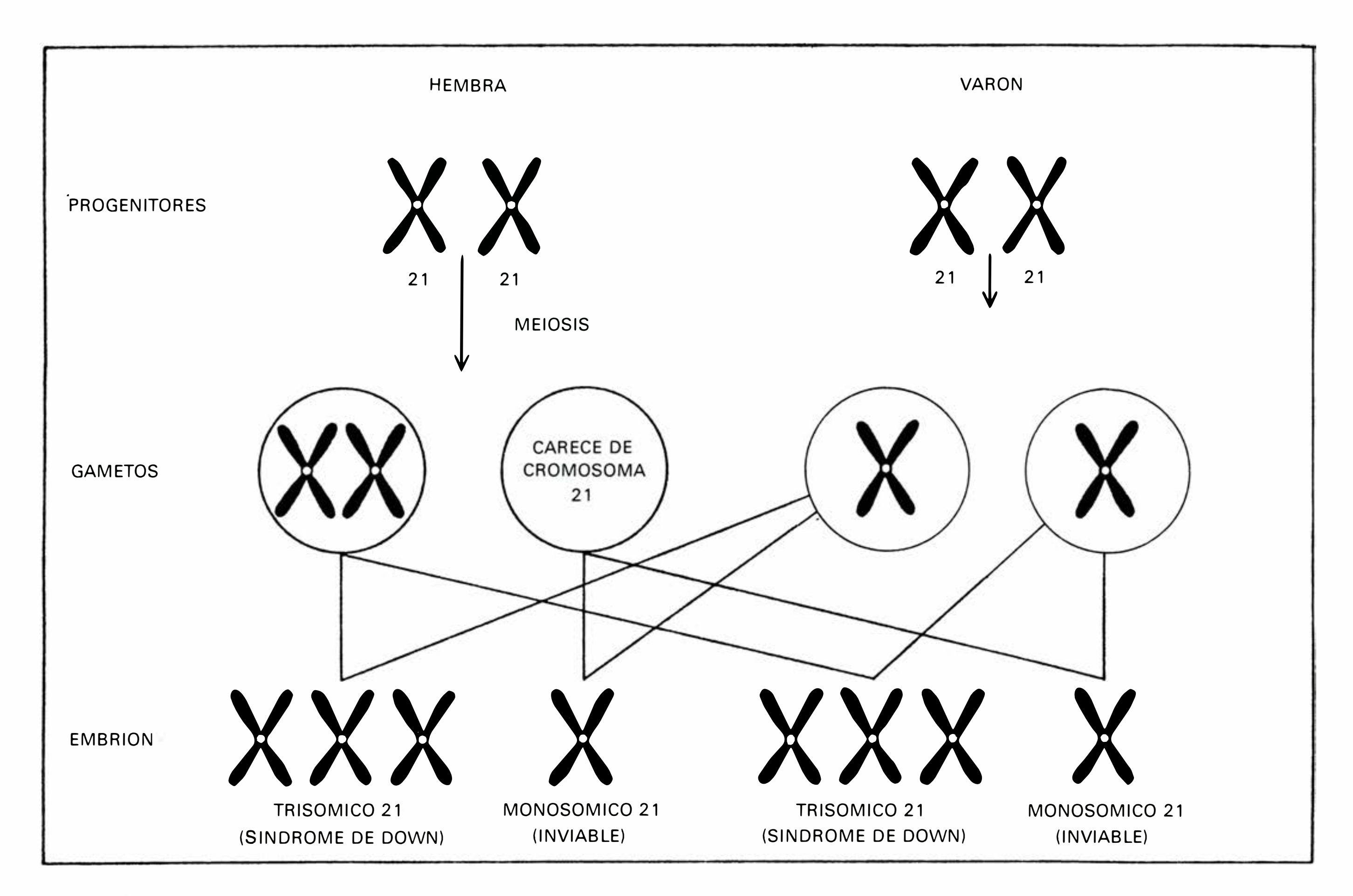
Las principales alteraciones de origen genético son las cromosomopatías y las enzimopatías.

Las alteraciones cromosómicas pueden producirse en los autosomas y en los gonosomas o cromosomas sexuales. La frecuencia de las cromosomopatías se reparte de la siguiente forma: en los recién nacidos, las alteraciones de los gonosomas es del 2,2 por mil y las de los autosomas es del 1,1 por mil; en los deficientes mentales, las cromosomopatías gonosómicas constituyen el 5,5 por mil y las autosómicas el 85 por mil; en los malformados tenemos cifras globales del 43 por mil y del 200 por mil en abortos. Dentro de las cromosomopatías autosómicas, y en cuanto se refiere a las alteraciones por reducción o aumento en el número de cromosomas, la mayoría de las monosomías (reducción) son letales. En cuanto a las trisomías (aumento), las que afectan al par 18 suelen conducir a la muerte antes de los 3 meses de vida; las del par 13 conllevan con frecuencia malformaciones.

Las más frecuentes son las del par 21: la trisomía 21, llamada síndrome de Down o mongolismo. Esta constituye la cromosopatía de mayor incidencia, lo que pone claramente de manifiesto su mayor grado comparativo de compatibilidad con las condiciones propias de la vida. A este respecto, debe precisarse que sólo uno de cada cinco fetos afectos de trisomía 21 llega a nacer, lo que supone que en el 80 por ciento de los casos se produce el aborto espontáneo. Buena parte de los casos se producen en las madres añosas, como muestran las estadísticas al respecto: la frecuencia del síndrome de Down por mil nacidos vivos es de 0,3 cuando la edad de la madre se halla entre los 20-29 años, 1,7 cuando oscila entre los 30-34 años, 3,5 si está entre los 35-39 años, 14,2 si entre los 40-44 años y 26,3 a partir de los 45 años.

La trisomía 21 se produce principalmente por no disyunción de los cromosomas durante la meiosis o durante la mitosis. Sólo está motivada por factores hereditarios en caso de traslocación, es decir, cuando un segmento cromosómico pasa a otra parte distinta de un cromosoma homólogo o heterólogo, durante el proceso de división celular. La frecuencia del síndrome de Down por no disyunción (47 cromosomas) es de un caso por cada 550 niños nacidos vivos, aproximadamente. El cariotipo normal consta de 23 pares cromosómicos.

Una cosa debe quedar bien clara: existe un extraordinario gradiente de gravedad dentro del síndrome de Down. Algunos niños presentan un cociente intelectual muy bajo y las posibilidades de su educación e incluso entrenamiento son muy escasas. En cambio, otros niños llegan a alcanzar un grado de prepara-



EL MONGOLISMO o síndrome de Down se produce cuando los cromosomas maternos no se separan durante la meiosis, de tal modo que algunos óvulos presentan, en lugar de uno, dos cromosomas del par 21. Cuando estos óvulos con un cromosoma más por no disyunción son fertilizados, el embrión se desarrolla anormalmente. Las trisomías 21 son viables en una proporción aproximada del 20 por ciento. En cambio, las monosomías 21 son inviables, es decir, no pueden sobrevivir. La no disyunción suele observarse en los óvulos de mujeres de más de 36 años.

ción e integración social extraordinario.

Entre las alteraciones que afectan a los gonosomas destaca el síndrome de Klinefelter (XXY) y puede producirse por no disyunción durante la meiosis o durante la mitosis.

Las alteraciones autosómicas recesivas sólo se manifiestan fenotípicamente en los homocigotos; en las variantes dominantes basta la presencia del gen anormal para que se manifieste en el fenotipo la alteración; en las alteraciones ligadas al sexo, la anormalidad se halla presente en los gonosomas X o Y y depende, a su vez, de que sea dominante o recesiva. Una condición recesiva ligada al cromosoma X sólo se manifiesta en las mujeres homocigóticas. En los varones, sin embargo, un gen presente en el único cromosoma X se manifiesta siempre, ya que no depende del efecto moderador que puede ejercer el segundo cromosoma X propio del sexo femenino.

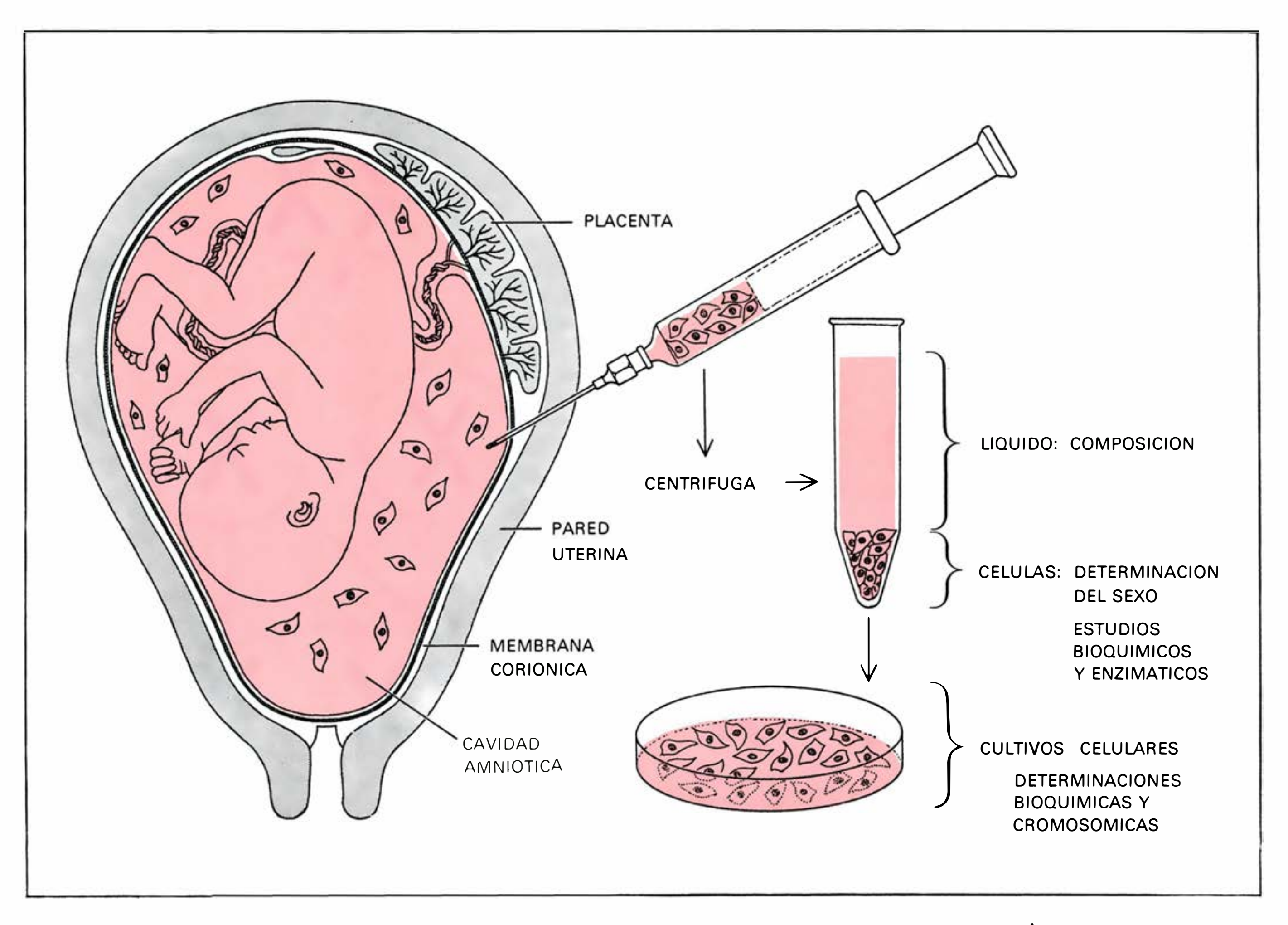
Todas estas alteraciones varían mucho en su expresividad fenotípica, según sean las características mutacionales del gen que se transmite. Pueden considerarse, en realidad, como heterocigotos cuya dotación no suple, evidentemente, la carencia del correspondiente *locus* genético deficitario.

Los cariotipos pueden determinarse también antenatalmente, es decir, puede saberse si el feto presenta alguna anormalidad cromosómica. Para ello, se obtiene líquido amniótico por amniocentesis, y se utiliza el propio líquido y las células que contiene, para diversas determinaciones bioquímicas y genéticas.

Enzimopatías

Todos los procesos de los seres vivos dependen de las reacciones enzimáticas, las cuales, en una forma u otra, actúan de una manera ordenada en una dirección determinada. La primera etapa de todos estos procesos es la *unión preferente de una molécula con otra,* lo que constituye una expresión del reconocimiento molecular. La disposición espacial de las proteínas, que depende a su vez de la estructura primaria, es decir, de la secuencia de los aminoácidos que la integran, es la base de su capacidad para reconocer estructuras específicas a través de unos puntos determinados de unión. Dichos puntos de unión poseen las características adecuadas para proporcionar la necesaria complementariedad geométrica y propiciar la unión, cuya solidez y especificidad dependerán del mayor o menor número de sistemas de engarce. El reconocimiento químico de una molécula por otra constituye una propiedad esencial de los enzimas, anticuerpos, hormonas, receptores, etcétera. El mismo "encaje espacial" constituye el fenómeno básico de todo el proceso genético: la complementariedad de la adenina con la timina y la citosina con la guanina (que constituyen las bases nitrogenadas del ADN) explica todos los mecanismos de reduplicación, transcripción y traducción de la información genética. Este reconocimiento químico, esta complementariedad, constituye la "inteligencia" de los fenómenos biológicos. Gracias a esta atracción química existe la especificidad que caracteriza al ser vivo.

Hoy conocemos ya con gran detalle los cambios secuenciales que se producen en las cadenas polipeptídicas integrantes de diversas proteínas a lo largo del *proceso evolutivo.* Las alteraciones secuenciales en las proteínas corresponden



OBTENCION DE LIQUIDO AMNIOTICO por amniocentesis: se obtiene una pequeña muestra del líquido que rodea al feto mediante una punción en la cavidad amniótica, cuidadosamente realizada para no dañar a la placenta ni al feto. El líquido amniótico, procedente principalmente de la orina y otras secreciones fetales, contiene células de origen fetal. Después de centrifugar, se realizan diversos análisis tanto en el líquido amniótico como en las células. El momento óptimo para practicar la amniocentesis se sitúa hacia la semana 16 de gestación.

a alteraciones secuenciales en los tripletes codificantes de los genes. Teóricamente, todas las mutaciones son posibles, si bien algunas de ellas afectan a unos aminoácidos "intangibles", ya que por hallarse directamente implicados en la recepción de los sustratos, coenzimas o cofactores, conducen a la síntesis de enzimas totalmente inhábiles para el ejercicio de su función. No es que no se produzcan enzimas, en términos generales, sino que se producen otras formas de enzimas, que poseen una actividad mayor, menor o nula, según sea su disposición espacial. Frente a los aminoácidos situados en posiciones invariables, otros son variables e incluso hipervariables. Sin embargo, conviene aclarar que, especialmente en las posiciones "variables", decir que pueden producirse muchas sustituciones no significa que pueden producirse todas las sustituciones.

La "lesión molecular" puede afectar únicamente a una de las unidades polipeptídicas que integran la proteína activa. Como sabemos, las formas menos regulables son las que caracterizan las situaciones de emergencia, tales como el periodo embrionario y el neonatal. El mantenimiento de las "formas embrionarias", por un defecto en la regulación de su biosíntesis, puede conducir al desarrollo de alteraciones de mayor o menor gravedad, como sucede en el caso de las talasemias, anemias de tipo hemolítico.

Las enzimopatías son en realidad cromosomopatías, a nivel molecular. Depende del nivel de observación. Como vemos, todo estado patológico es el resultado de una alteración de las actividades enzimáticas: de su síntesis o de su actividad, al modificarse los requisitos necesarios para su regulación o las condiciones microecológicas en que la actividad se desenvuelve.

Sólo es visible lo que no es letal. Además, según la funcionalidad del enzima resultante, se configura un amplio intervalo de lo normal del mismo modo que existe un amplio intervalo de lo patológico. Como hemos indicado, ambos intervalos dependen, a su vez, de que los enzimas afectados sean o no cruciales para el metabolismo ya que, lógicamente, los umbrales de tolerancia dependerán de los niveles de actividad normalmente necesarios para el adecuado funcionamiento de las distintas vías metabólicas.

Otro dato que es necesario considerar es el momento *de su requerimiento.* Algunos enzimas son indispensables en cualquier periodo del desarrollo; otros se precisan únicamente a partir de cierta etapa. Así, si una actividad enzimática determinada es necesaria para el desarrollo intrauterino, conducirá a una alteración prenatal, siempre que no se trate de una actividad crucial o que la lesión molecular no sea suficientemente grave. En cambio, el desarrollo posnatal de un estado patológico se producirá cuando la actividad implicada sólo sea necesaria a partir del momento

en que el nuevo ser adquiera su autonomía, bien porque hasta este momento el metabolismo no precisa dicha actividad o bien porque queda suplida por los resortes maternos.

La importancia del entorno

Citemos, por último, la extraordinaria *importancia del entorno.* Según sea eventual o normal la presencia de una componente del mismo, que exige una determinada actividad enzimática, será mayor o menor la gravedad de la situación patológica, que refleja la inhabilidad –asimismo, mayor o menor– de la proteína enzimática para el ejercicio de la función que le es propia.

La incidencia de las enzimopatías o metabolopatías congénitas es baja, verbigracia: cistinuria, 1 por 7000, fenilcetonuria 1 por 10.000, histidinemia 1 por 24.000, enfermedad de Hartnup 1 por 26.000, galactosemia 1 por 78.000 y homocistinuria 1 por 200.000. En algunos casos puede prevenirse totalmente la manifestación de la enfermedad mediante un diagnóstico precoz y el oportuno tratamiento.

El número de "errores congénitos del metabolismo" aumenta sin cesar. Se conocen enzimopatías del metabolismo de los aminoácidos, glúcidos y lípidos; de las bases púricas y pirimidínicas; del metabolismo de las vitaminas y de los oligoelementos; de las hormonas... Estas alteraciones conllevan normalmente un profundo retraso mental. A veces, se ha llegado a conocer con precisión la "razón molecular" de las alteraciones patológicas que caracterizan a estas enfermedades, como ha ocurrido con las homocistinurias, que son enzimopatías del metabolismo de un aminoácido azufrado, la metionina.

En nuestro laboratorio, Valdivieso y colaboradores han estudiado las repercusiones metabólicas a que puede conducir la interferencia de las actividades enzimáticas de procesos fundamentales para el cerebro durante el periodo neonatal, tales como la cetogénesis. Este proceso, plenamente fisiológico en el recién nacido, puede ser inhibido por la fenilalanina o sus derivados en el caso de enfermos fenilcetonúricos, de tal modo que se inhiba la mielinización, lo que explicaría la lesión cerebral propia de esta enzimopatía.

Con animales de experimentación con enzimopatías simuladas, los autores han investigado qué efecto tienen, sobre el metabolismo de los cuerpos cetónicos, ciertos productos derivados de la fenilalanina, que llegan a producirse en el quehacer fisiológico normal, pero que se acumulan en el cerebro del paciente fenilcetonúrico. Es posible que la inhibición de la mielogénesis que dichas sustancias pueden ejercer (en el contexto enzimático *propio* del neonato) contribuyan a explicar la etiología del daño cerebral. Ciertamente, es bien conocido que en las glucogénesis, la carencia de determinados enzimas propios de la glucogenogénesis y de la glucogenolisis, producen alteraciones substanciales en la estructura del polímero que se acumula y, por tanto, en la posibilidad de su utilización.

Algo parecido sucede en el caso de las lipidosis: la producción de estructuras lipídicas inadecuadas para el ejercicio de las funciones propias de los lípidos complejos conduce a sistemas imperfectos en la composición o protección del sistema nervioso central, de tal manera que la etiología de la enfermedad es, en estos casos, más fácilmente deducible.

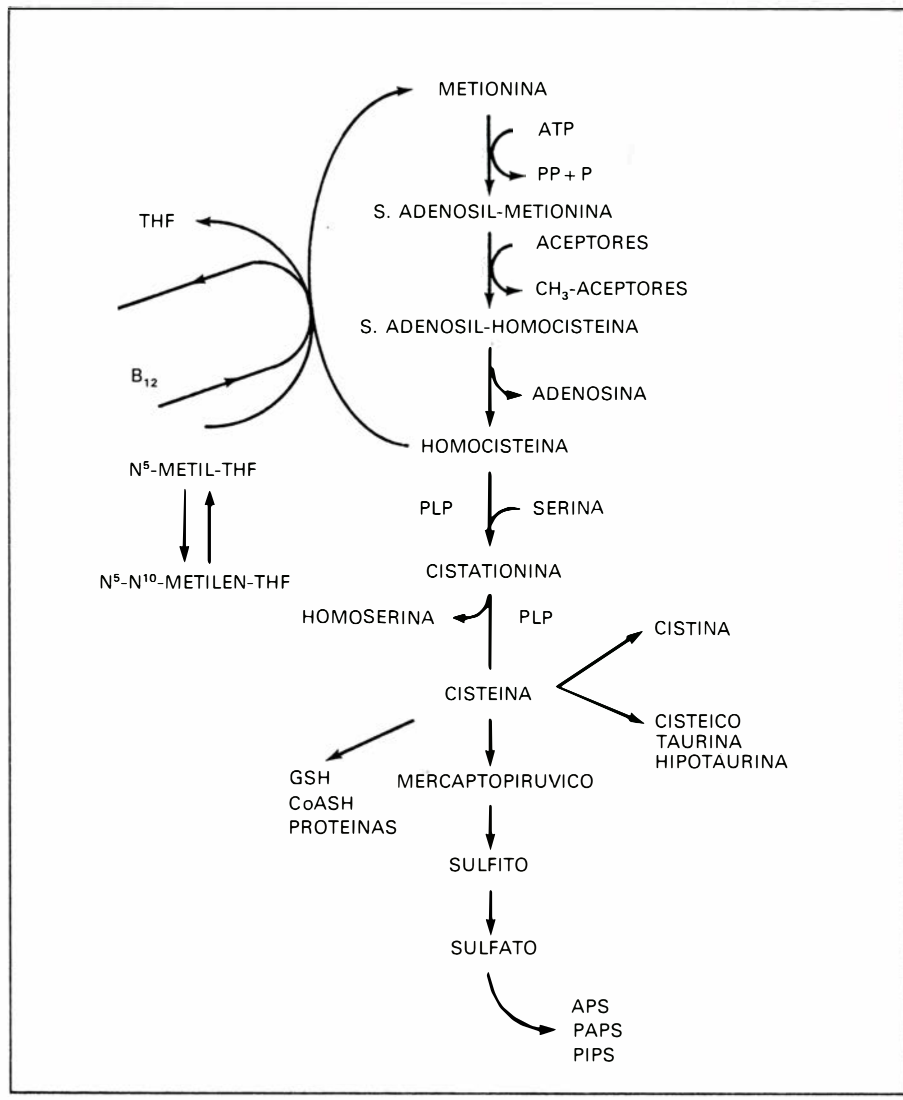
No sucede lo mismo en las aminoacidopatías o en las enzimopatías que afectan al metabolismo de las bases púricas y pirimidínicas. En estos casos, es el acúmulo de una sustancia normal o de sus derivados, menos habituales, el que produce en un momento determinado –el del máximo desarrollo del cerebro– interferencias metabólicas en el periodo neonatal. Puede aducirse que, por una menor capacidad de selección en la barrera hemato-encefálica, tengan acceso al cerebro sustancias que, más adelante, verían impedido su paso al mismo. Pero, alternativa o simultáneamente, puede pensarse que el acúmulo de estas sustancias *inhibe el funcionamiento de sistemas específicos del periodo neonatal.* Así, el metabolismo de los cuerpos cetónicos, tan importante en dicho periodo, en el que es plenamente fisiológico, adquiere ulteriormente características eventuales que, si se agudizan, condicionan situaciones claramente patológicas.

La posible interferencia del proceso que conduce a la síntesis de las sustancias que forman parte de las mielinas cobra mayor sentido si consideramos que en el cerebro del neonato son plenamente funcionales durante las primeras semanas o meses de su vida unos procesos bioquímicos que, más tarde, actuarán exclusivamente en situaciones de emergencia o se situarán en el terreno de lo patológico. Tal como hemos comentado, este es el caso del metabolismo de los cuerpos cetónicos. que se producen en el hígado en grandes cantidades en situaciones en las que el aporte de glucosa es muy precario y se consumen principalmente grasas. Pero que no pueden ser utilizados en el hígado, por carecer de un enzima, la 3-oxoácido transferasa, que se halla en otros tejidos y, especialmente, en el cerebro.

Del mismo modo que existe un metabolismo específico del feto, *existe un metabolismo específico del neonato.* En consecuencia, sólo deben utilizarse los datos procedentes de la experimentación en las diversas etapas del desarrollo intrauterino y posnatal, como puntos de referencia para la interpretación correcta de los resultados obtenidos en la investigación de alteraciones producidas a lo largo de este periodo.

Las homocistinurias

La homocistinuria es un buen ejemplo para el estudio de las alternativas metabólicas que pueden conducir al acúmulo de un mismo metabolito intermediario, así como del *extraordinario número de variantes que cada tipo de enzimopatía posee.* Además, el metabolismo de la homocistinuria permite establecer hipótesis muy razonables sobre el origen molecular de la sintomatología que presentan los enfermos afectos de dichas enzimopatías. Muy brevemente, el acúmulo de homocistina puede producirse: *a)* por déficit del enzima cistationina sintetasa, enzimopatía que impide la síntesis de cisteína, pero que permite e incluso favorece la conversión de homocisteína en metionina y, en consecuencia, las reaccciones de transmetilación subsiguientes, de extraordinaria importancia en el organismo y, de manera concreta, para la biosíntesis de la colina y de la adrenalina; *b)* por disfunción del sistema de conversión de homocisteína en metionina, disfunción que a su vez presenta diversas modalidades, ya que puede producirse por un fallo en el sistema reductásico que convierte el ácido 5-10 metilentetrahidrofólico en 5-metiltetrahidrofólico o en la deficiente absorción, incorporación celular o metabolismo degradativo de la vitamina B_{12}. En todos estos casos, se acumula homocisteína en menor cantidad, y la sintomatología principal y secundaria es de menor relieve. En la homocistinuria "grave" se produce un considerable retraso mental acompañado de ectopia con dislocación lenticular progresiva, falleciendo normalmente los enfermos por obstrucción circulatoria. Es fácil imaginar la conexión existente entre estas alteraciones a nivel



LAS ENZIMOPATIAS CONOCIDAS en el metabolismo de los aminoácidos azufrados son numerosas. Las más frecuentes cursan con acúmulo de homocistina, procedente de la homocisteína que se origina bien sea por incapacidad para catalizar la transformación de homocisteína en cistationina (déficit en cistationina sintetasa) o por fallo del sistema que convierte a la homocisteína en metionina, de tanta importancia en la bioquímica de la transmetilación. La distinta respuesta que presentan los pacientes a la administración de factores coenzimáticos tales como el piridoxal.5. fosfato (PLP) pone de manifiesto el extraordinario número de variantes de cada enzimopatía, según sea la "lesión molecular" resultante: la sustitución de un aminoácido por otro (consecuencia de una mutación en los genes) puede alterar con mayor o menor intensidad la disposición espacial de la proteína enzimática y, en consecuencia, su capacidad funcional. La interferencia, a cualquier nivel, del esquema metabólico representado conduce a una menor producción de sulfato activo, lo que permite explicar satisfactoriamente el origen molecular de algunos síntomas característicos que presentan los enfermos afectos de estas alteraciones (dislocación lenticular, accidentes circulatorios por obstrucción, etc.) (T.H.F.: tetrahidrofólico; B_{12}: cobalamina o vitamina B_{12}; GSH: glutation; APS: adenosín fosfosulfato; PAPS: fosfoadenosín fosfosulfato; PIPS: fosfoinosín fosfosulfato).

molecular y una deficiente síntesis de los ácidos condroitín y mucoitín sulfúrico, así como de la heparina, que requieren para su formación el adecuado aporte de "sulfato activo", que no tiene lugar debido a que la ruta principal que conduce a la formación de PAPS se halla bloqueada. En favor de esta hipótesis puede citarse el hecho de que la inhibición de la sulfito oxidasa produce en los pacientes una dislocación lenticular de muy precoz desarrollo.

El hecho de que tanto la homocistinuria como la cistationuria presenten un extraordinario abanico de respuestas frente a la administración de vitamina B_6 pone de manifiesto la extraordinaria diversidad de modalidades que, en virtud de una modificación en la secuencia de aminoácidos compatible con la vida, existen en cada tipo de enzimopatía. Lo mismo puede decirse en el caso de la fenilcetonuria, así como en la respuesta a la administración de vitamina B_1 en la enfermedad de la orina con olor a jarabe de arce.

La lesión molecular se ha producido, en el último caso, en la estructura de la apo-descarboxilasa específica de cada uno de los aminoácidos de cadena ramificada. Dada la importancia de la descarboxilación del piruvato y del α-cetoglutarato en el metabolismo, es difícil imaginar lesiones moleculares "profundas" en estos enzimas, habida cuenta del angular papel que desempeñan en el metabolismo.

Hipotiroidismo

Un ejemplo de alteración endocrina congénita lo constituye el hipotiroidismo. El hipotiroidismo congénito, con una frecuencia de uno a tres casos por seis mil recién nacidos, da lugar a un retraso mental grave que puede ser irreversible si el tratamiento de substitución con hormonas tiroideas no se inicia precozmente.

Aparte de la deficiencia intensa en hormonas tiroideas, hay otros casos (con una frecuencia de uno en 500) en que los niveles de tiroxina (T4) sérica son significativamente más bajos de los normales, sin sintomatología clínica aparente. Como han puntualizado Morreale y Escobar, "el problema estriba en que no se sabe si estos niños, estudiados al cabo de los años tendrán, como grupo, un desarrollo mental inferior al de los niños con niveles normales de T4 en el periodo perinatal". ¿Hay acaso un estado de hipotiroidismo neonatal, posiblemente temporal, parecido a la hipertiroxidemia neonatal? Y si lo hay, ¿necesitarían tales niños una terapéutica de sustitución?

No sabemos todavía si los niños con grado leve o parcial de hipotiroidismo están más expuestos al riesgo de retraso mental, de no ser tratados, que los otros. La lección aprendida de las formas atípicas de hiperfenilalaninemia debería haberse aprendido bien.

El tratamiento con hormonas tiroideas tiene que iniciarse antes de los tres meses de vida. Smith y colaboradores encontraron que el 45 por ciento de los hipotiroidismos congénitos tratados antes de los seis meses alcanzaban un coeficiente intelectual de 90 o más, mientras que sólo un 10 por ciento de los tratados a partir de los seis meses alcanzaban un coeficiente intelectual relativamente normal. Antes de los tres meses, el porcentaje asciende al 75 por ciento.

Sufrimiento fetal durante el parto. Prematuros.

Vistas las principales alteraciones hereditarias pasamos ahora a revisar algunos ejemplos de alteraciones que son adquiridas.

A partir de la década de los cincuenta se han hecho patentes las secuelas permanentes y complicaciones del embarazo, tales como parálisis cerebral, deficiencia mental, epilepsia, dislexia, etcétera, lo que ha corroborado las previsiones establecidas hace más de doscientos años por Smellie y reiteradas en 1861 por Little en "Transactions of the London Obstetrical Society". Poco a poco se van estableciendo interesantes correlaciones entre alteraciones neurológicas y prematuridad, o hemorragias prenatales, toxemia, placenta previa, prolapso del cordón, etcétera. Al mismo tiempo se inicia la búsqueda de los parámetros que pueden reflejar el bienestar fetal sin necesidad de instalaciones complicadas y costosas. En honor de su autor, que hace una veintena de años publicó los resultados de sus trabajos sobre la "evaluación del neonato", la vitalidad del recién nacido se expresa en unidades "Apgar".

Si a esta práctica se uniera la realización de tres visitas al médico durante el embarazo, de carácter obligatorio y en las que se hiciera constar, en un carnet "de la gestante y del recién nacido", las constantes más elementales; si el ritmo cardíaco fetal fuera seguido gráficamente y, cuando fuera necesario, pudiera analizarse el equilibrio ácido-base y se dispusiera de una incubadora bien equipada para emergencias, podría evitarse un elevado porcentaje de los niños afectos actualmente de minusvalías.

Los fenómenos de asfixia se presentan en el 54 por mil de los nacidos vivos de ellos, siete quedarán con una lesión cerebral permanente. El gran problema consiste en identificar a los que corren tan grave riesgo.

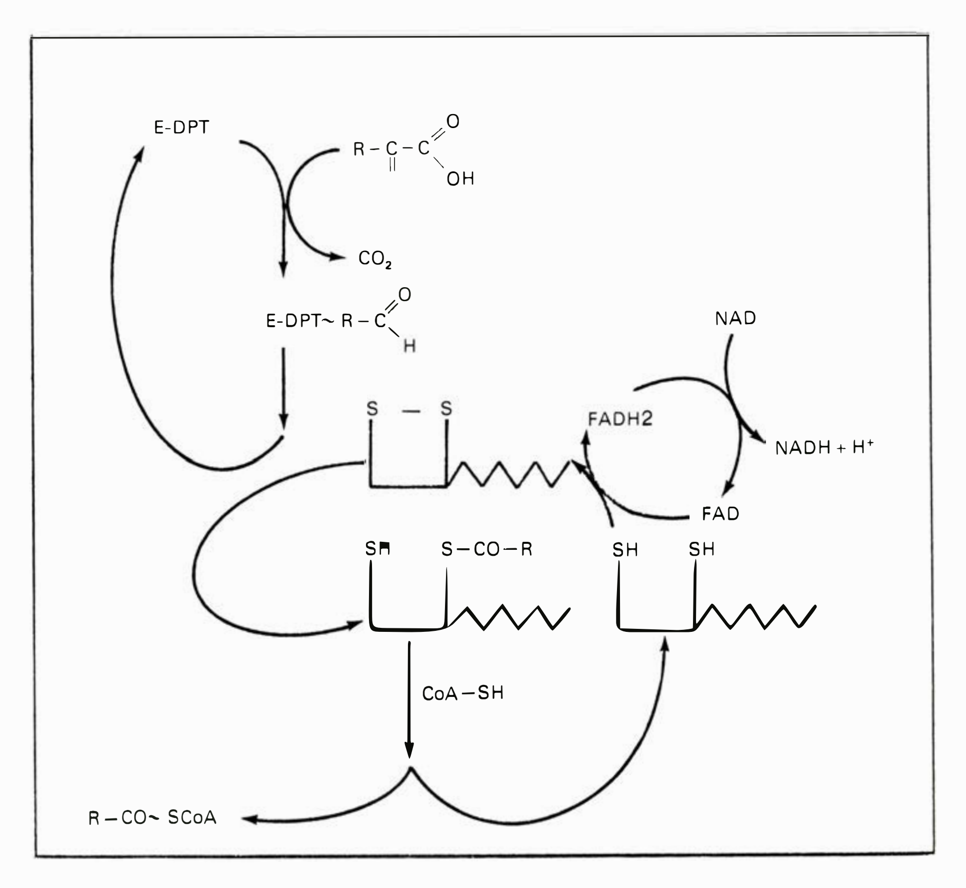
La presentación anormal del feto en el momento del parto, un reducido tamaño de la pelvis materna, la placenta previa y prolapso del cordón umbilical son algunas de las principales causas de alteraciones neurológicas resultantes del parto. Además de las pruebas de Apgar, debería constatarse si ha habido dificultades respiratorias en los primeros minutos de vida, infecciones neonatales, sintomatología ostensible a la supervisión neonatológica, existencia de hiperbilirrubinemia, hipoglucemia o hipotermia.

Superados en buena parte los traumas físicos, deben evitarse los traumas "fisiológicos", especialmente cuando es presumible –dadas las características del proceso gestacional, prematuridad, etcétera– que el feto se halla precariamente preparado para sufrir la auténtica convulsión bioquímica que representa el parto. Sería muy conveniente que pudiera objetivarse el estado de madurez que presenta el feto para hacer frente al tránsito por las vías naturales y a las exigencias (sobrepresión, tiempo transcurrido entre el cese del aporte placentario y la ventilación, etc.) que éste comporta.

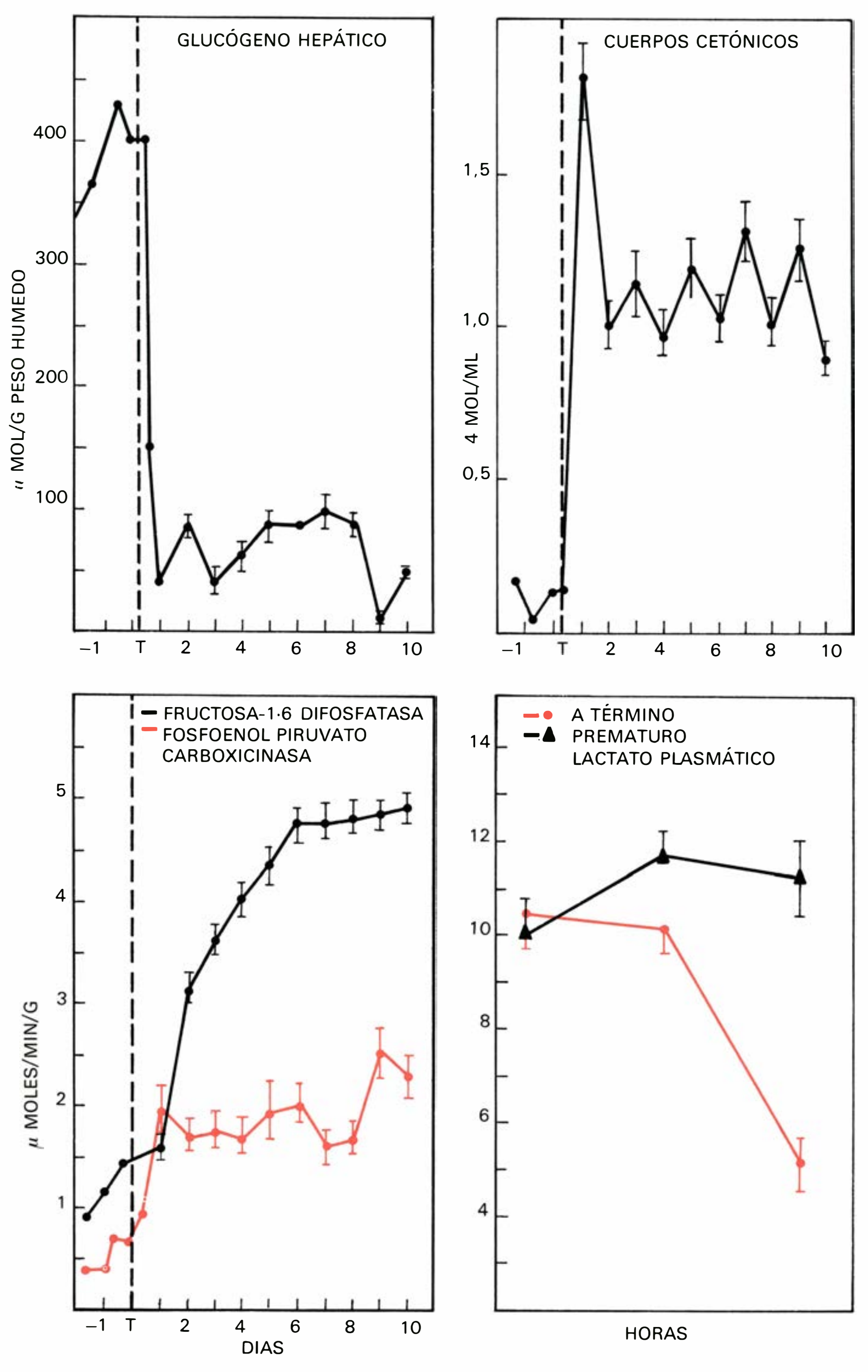
A este respecto, resultan bastante significativos los experimentos que han llevado a cabo Medina y colaboradores, en nuestro laboratorio, con ratas extraídas por cesárea al término del periodo de gestación y 24 horas antes. Se observa, por ejemplo, que la capacidad para utilizar el lactato, inmediata en el caso del animal nacido a término, sufre una demora sensible en el prematuro, que aparece, por tanto, como más sensible –y, en consecuencia, vulnerable– a las condiciones que conlleva el parto. Si se consiguiera disponer de unos índices que reflejaran el grado de prematuridad real o, lo que es lo mismo, la madurez funcional del feto, bastaría con aplicar la cesárea en aquellos casos en que así fuera aconsejable para que se redujera sustancialmente la incidencia de casos de subnormalidad de etiología o causalidad obstétrica.

La resistencia a la situación hipóxica decrece en cuestión de horas después del nacimiento. En otras palabras, la capacidad para obtener energía exclusivamente por vía glucolítica se desarrollará sólo durante un periodo de varios minutos en el momento del nacimiento para utilizar a continuación, en una situación de hipoglucemia característica del postparto, otros metabolitos como combustible biológico, verbigracia, los cuerpos cetónicos y el mismo lactato acumulado, que se degradaría ahora por vía oxidativa.

La inducción de los sistemas gluconeogénicos se produce después del parto, muy rápidamente, pero no se hallan preparados los enzimas correspondientes en el momento del nacimiento. Como vemos, es el propio lactato acumulado en la glucolisis intraparto el que se oxida rápidamente, el mismo tiempo que los cuerpos cetónicos se metabolizan intensamente.



LAS REACCIONES DE DESCARBOXILACION OXIDATIVA requieren la participación de un complejo sistema enzimático en el que intervienen el ácido lipoico y el coenzima A. La dihidrolipoil transacetilasa transfiere el grupo acilo al coenzima A, al mismo tiempo que el ácido lipoico se transforma en dihidrolipoico y se reoxida a la forma disulfuro por acción de otro enzima, la dihidrolipoil deshidrogenasa, cuyo coenzima es el FAD. Dado el general e importante papel de estos enzimas, sólo pueden preverse alteraciones leves en su estructura, ya que las que afectaran gravemente su capacidad funcional no serían viables. (DPT: difosfato de tiamina; E: enzima; FAD: flavín.adenín.dinucleótido; NAD: nicotinamida.adenín.dinucleótido; por fin, CoASH: coenzima A).



EVOLUCION DE ENZIMAS GLUCONEOGENICOS, del glucógeno, lactato y cuerpos cetónicos en ratas, en las cuales se observa el súbito metabolismo de las reservas hidrocarbonadas y su conversión en lactato, al mismo tiempo que los cuerpos cetónicos ocupan un protagonismo evidente, en un momento inicial en el que los enzimas de la gluconeogenésis no se hallan todavía "preparados". Esta falta de maduración se pone de manifiesto en los fetos extraídos por cesárea un día antes de cumplirse el periodo normal de gestación: son incapaces de metabolizar el lactato, al mismo ritmo, al menos, que lo hace la rata extraída por cesárea a término. (De José M.ª Medina y colab.).

Se considera que aproximadamente un 70 por ciento de la mortalidad perinatal se halla asociada con bajo peso al nacer. Por otra parte, también un tercio, como mínimo, de las parálisis cerebrales se presentan en estos prematuros "reales". Se ha conseguido, en aquellos lugares en donde se han llevado a la práctica elementales normas de atención al embarazo y al neonato una sensible reducción de los accidentes perinatales. La parálisis cerebral ha descendido, por ejemplo, del 2,5 por mil en el periodo 1948-1953 al 1,6 por mil en 1953-1958 y al 0,9 por mil en el intervalo 1958-1962 (datos de Bristol, Inglaterra). En contraste, en algunos países (como Nigeria) el número de casos de parálisis cerebral se eleva al 10 por 1000. El factor etiológico predominante parecen ser ictericias del neonato (el 70 por ciento de los casos sin causa conocida) e infecciones posnatales, tales como meningitis y encefalitis.

El sufrimiento fetal crónico, que tiene lugar durante el embarazo, puede seguirse actualmente mediante la valoración de lactógeno placentario, que refleja el funcionamiento de la placenta; de estradiol, que pone de manifiesto la funcionalidad de las cápsulas suprarrenales del feto también algunos enzimas, como los fosfatasa alcalina termoestable, pueden servir de índice para establecer el bienestar fetal.

El sufrimiento fetal agudo es el que se produce en el momento del parto, siendo de naturaleza principalmente respiratoria. Mediante ultrasonido puede seguirse el número de movimientos torácicos por minuto que presenta el feto, como parámetro de vitalidad del mismo.

Incompatibilidad fetomaterna. Rubéola

Desde que Lansteiner y Wiener comunicaron en 1941 la existencia del antígeno Rh y la posibilidad de formación en la madre de isoanticuerpos y, por ende, la patogénesis de la eritrobalastosis fetal, hasta que, siete años más tarde, se propuso por Diamond la exanguinotransfusión son numerosos los casos de ictericia grave del recién nacido que pueden evitarse cuando se practica el análisis de eritrosedimentación pertinente. Sin embargo, la transfusión del neonato pronto podría ser reemplazada por un sencillo tratamiento con inmunoglobulinas, gracias a los trabajos de Pollock, Gorman y Freda.

En caso de incompatibilidad fetomaterna no tratada, se deriva una encefalopatía ictérica, producida por un acúmulo de bilirrubina. La hiperbilirrubinemia puede tener múltiples orígenes, genéticos o adquiridos; se produce por: incompatibilidad Rh, anemias hemolíticas por intoxicaciones, defecto congénito en glucuronil transferasa, gran prematuridad que conlleva inmadurez hepática, y otros. La bilirrubina ejerce un efecto muy tóxico sobre el metabolismo celular, siendo entonces aconsejable practicar la exanguinotransfusión.

En España se producen de 200 a 300 casos de malformaciones congénitas debidas a padecer la rubéola durante el embarazo. Con las correspondientes inmunoglobulinas pueden evitarse actualmente las graves consecuencias del

"morbo hemolítico del neonato" y de la rubéola. Gracias a este proceso inmunitario, al salvar uno la incompatibilidad fetomaterna propia del Rh y el otro la infección vírica de la rubéola en el periodo embriogénico, han podido neutralizarse dos de las principales causas de subnormalidad evitable.

Es innecesario insistir en la importancia que la poliomielitis, la difteria, el tétanos y la tos ferina pueden tener como causas de subnormalidad "adquirida".

Malnutrición

Existe una considerable controversia sobre la influencia de la malnutrición en el desarrollo mental del niño. Es evidente que no pueden extrapolarse de manera taxativa los resultados hallados en los animales de experimentación al caso del hombre. También es obvio que es más importante la alimentación de la madre durante la gestación que durante la lactancia. Sin embargo, durante todo el proceso de maduración del cerebro, la malnutrición puede acarrear secuelas permanentes, más o menos graves según sea: *a)* la dotación genética de cada individuo, de tal modo que le permita hacer frente con mayor o menor capacidad de manejo a la precaria ingesta proteica; *b)* la restricción proteica que caracterice a la dieta normal de la madre durante la gestación y la lactancia, ya que es previsible que en muchos casos la ingesta proteica sea sensiblemente inferior a la que se utiliza en las pruebas de subnutrición en animales de experimentación. Nosotros hemos hallado en el norte de la provincia de Granada, en lugares en donde la alimentación es muy uniforme y de contenido escasamente proteico, una disfunción renal caracterizada por aminoaciduria generalizada en los primeros meses de vida, con clásica sintomatología de raquitismo a los tres meses, en un porcentaje superior al 30 por ciento de los niños analizados.

En cualquier caso, la repercusión de la malnutrición en el cociente intelectual, en el comportamiento y en la capacidad de aprendizaje, tendrá todavía un gradiente mucho más amplio que otras causas de alteración molecular cerebral, ya que se trata de apreciaciones cualitativas de difícil distinción, poniéndose de manifiesto de manera evidente y clara únicamente aquellas que son muy graves bien sea por una dieta extraordinariamente desviada o por unas capacidades genéticas caracterizadas por un umbral de tolerancia muy exiguo en lo que a la ingesta proteica se refiere.

La paradoja de Stein en estadística

La predicción óptima de un suceso futuro se elabora promediando los sucesos pasados. La paradoja de Stein determina las condiciones de existencia de estimadores que son mejores que la media aritmética

Bradley Efron y Carl Morris

Hay ocasiones en que un resultado matemático, cuya demostración es a todas luces válida, está en abierta contradicción con el sentir general. En 1955, Charles Stein, de la Universidad de Stanford, descubrió una de estas paradojas en estadística. Su resultado socavó siglo y medio de trabajos en teoría de la predicción, que se remontan hasta Karl Friedrich Gauss y Adrien Marie Legendre. Tras largo tiempo de resistencia a las ideas de Stein, salpicado por discusiones frecuentes y a veces ásperas, parece que la sensación paradójica ha disminuido, y dichas ideas comienzan a introducirse en estadística pura y aplicada.

La paradoja de Stein se refiere al empleo de promedios observados al objeto de estimar el valor de magnitudes no observadas. Entre las operaciones de la estadística, la obtención de promedios ocupa, por orden de importancia, el segundo lugar, estando el primero reservado al simplicísimo acto de contar. De un jugador de béisbol que consiga siete tantos en 20 ensayos con el bate durante un partido, se dice que tiene una "habilidad" de 0,350. Al calcular este valor se efectúa una estimación de la verdadera habilidad del jugador, usando para ello el promedio de sus tantos pasados. Si se nos preguntara qué resultados obtendrá el jugador en sus próximos 100 ensayos con el bate, probablemente diríamos que 35 tantos más. En la teoría estadística tradicional se puede demostrar que ninguna otra regla de estimación es uniformemente superior al promedio de observaciones.

Lo paradójico del resultado de Stein es que a veces contradice esta ley elemental de teoría estadística. Si se tienen tres o más jugadores de béisbol, y se desea estimar la habilidad futura de cada uno de ellos, hay un procedimiento mejor que la pura extrapolación a partir de sus promedios personales. Y el calificativo "mejor" es aquí muy significativo. Un estadístico que aplique a este problema el método de Stein puede esperar predicciones más exactas de los respectivos promedios futuros, independientemente de cuáles sean las habilidades verdaderas de los jugadores.

El béisbol es un deporte en que se han elaborado cuidadosamente largas series estadísticas, y que proporciona material muy adecuado para exponer el funcionamiento del método de Stein. Como datos primarios tomaremos los promedios de 18 jugadores de liga de primera división, computados tras sus primeros 45 turnos de bate, durante la temporada de 1970. Estos jugadores fueron todos los que completaron 45 turnos el mismo día que se hizo la tabulación. Los promedios se calculan, sencillamente, dividiendo el número de tantos por el número de turnos de bate; siempre es un número comprendido entre 0 y 1. Cada promedio se representará por la letra y.

El primer paso para aplicar el método de Stein es calcular el promedio de los promedios. Evidentemente, este "gran promedio", que representaremos por $\bar{y}$ estará también comprendido entre 0 y 1. El proceso esencial del método de Stein consiste en "contraer" todos los promedios individuales hacia este gran promedio. Si la marca de un jugador es superior al gran promedio, se debe disminuir dicha marca; si no alcanza el gran promedio, se deberá aumentar el tanteo previsto para el jugador. Designaremos por z al valor contraído resultante para cada jugador. Este valor se llama estimador de James-Stein de la habilidad del jugador, en honor de Stein y de W. James, quienes, en 1961, propusieron una versión especialmente sencilla del método. La paradoja de Stein consiste, simplemente, en que los valores z, es decir, los estimadores de James-Stein, dan mejores estimaciones de la verdadera habilidad de los jugadores que sus promedios simples individuales.

El estimador James-Stein de cada jugador se halla mediante la siguiente fórmula: $z = \bar{y} + c\,(y - \bar{y})$. El término $(y - \bar{y})$ expresa la diferencia entre el promedio de cada jugador y el gran promedio. La fórmula expresa que el estimador de James-Stein, z, se diferencia del gran promedio en esta misma cantidad $(y - \bar{y})$ multiplicada por una constante, c. La constante c se llama "coeficiente de constricción". Si fuese igual a 1, la fórmula diría que el estimador de James-Stein de un jugador dado es idéntico al promedio simple de dicho jugador; es decir, que y es igual a z. El teorema de Stein enuncia que c siempre es menor que 1. Su valor real viene determinado por la colección de todos los promedios observados.

En el caso de los datos del béisbol, el gran promedio, $\bar{y}$, es 0,265, y el coeficiente de constricción, c, es 0,212. Sustituyendo estos valores en la fórmula, vemos que, para cada jugador, z es igual a $0{,}265 + 0{,}212\,(y - 0{,}265)$. En razón de que c es aproximadamente 0,2, cada promedio se "contraerá" alrededor de un 80 por ciento de su diferencia con el gran promedio, y la dispersión total de los promedios disminuirá también en aproximadamente un 80 por ciento.

A título de ejemplo, tomemos al fallecido Roberto Clemente, que era el bateador más destacado de las ligas de primera división de la época en que se tabularon nuestros datos estadísticos. Para Clemente, y es igual a 0,400, y z puede calcularse mediante la expresión $z = 0{,}265 + 0{,}212(0{,}400 - 0{,}265)$. Resulta 0,294. Dicho de otro modo, el teorema de Stein afirma que la verdadera habili-

ROBERTO CLEMENTE
FRANK ROBINSON
FRANK HOWARD
JAY JOHNSTONE
KEN BERRY
JIM SPENCER
DON KESSINGER
LUIS ALVARADO
RON SANTO
RON SWOBODA
DEL UNSER
BILLY WILLIAMS
GEORGE SCOTT
RICO PETROCELLI
ELLIE RODRIGUEZ
BERT CAMPANERIS
THURMAN MUNSON
MAX ALVIS

0 .050 .100 .150 .200 .250 .300 .350 .400 .450
HABILIDAD PROMEDIO

0 .005 .010 .015 .020 .025
ERROR CUADRATICO



LA HABILIDAD CON EL BATE de 18 jugadores de béisbol de primera división puede estimarse más exactamente por el método de Charles Stein y W. James que mediante sus promedios individuales. Los promedios usados como estimadores se han calculado tras 45 turnos de bate para cada jugador, durante la temporada de 1970. La verdadera habilidad de un jugador es una cantidad inobservable, que puede sustituirse con buena aproximación por su promedio a largo plazo. La verdadera destreza se representa aquí mediante el promedio mantenido durante el resto de la temporada de 1970. Para 16 de los 18 jugadores, el promedio inicial predice peor la habilidad con el bate que el estimador de James-Stein. El conjunto de estimadores de James-Stein muestra también el menor error cuadrático total.

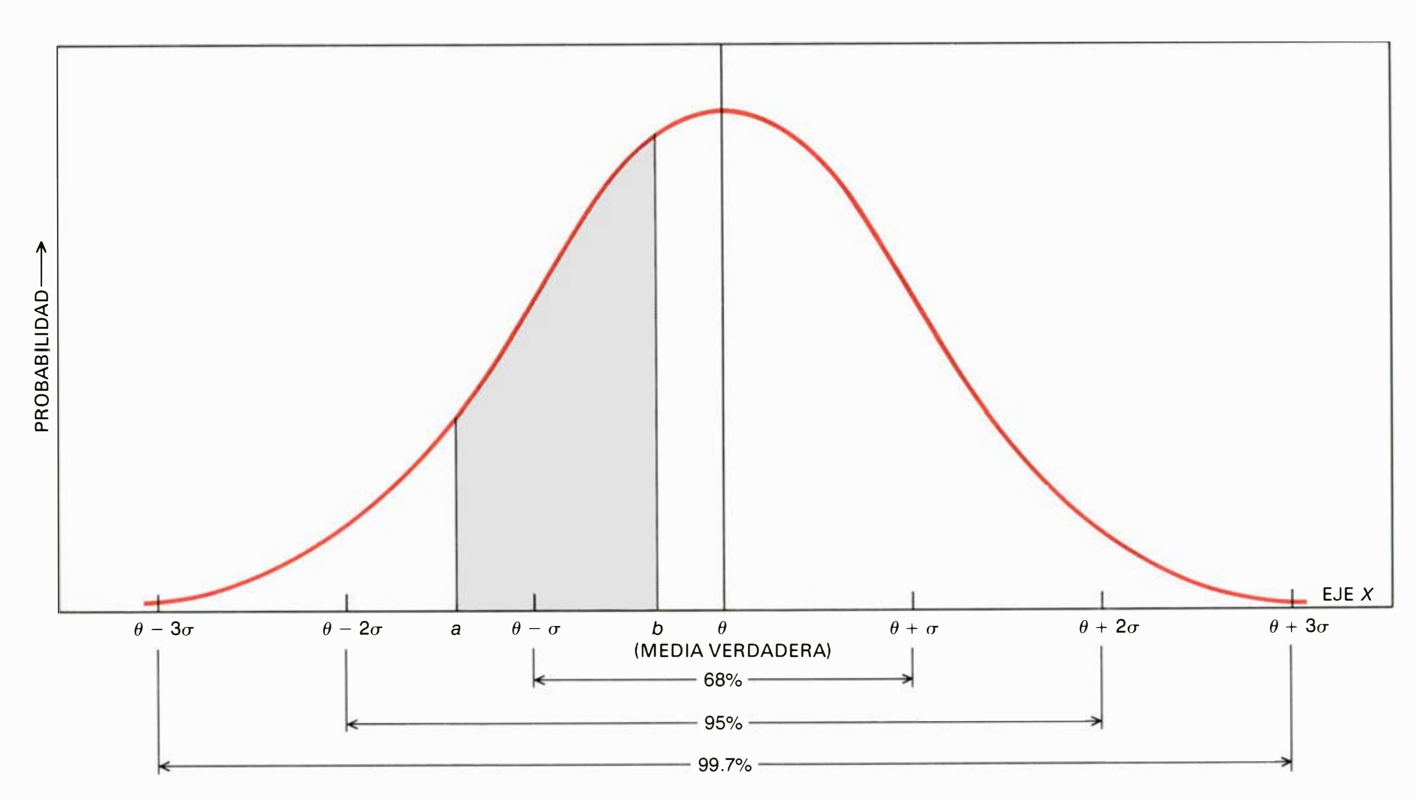
dad de Clemente puede estimarse con mayor precisión en 0,294 que en 0,400. Thurman Munson, que sufrió un brusco hundimiento al comienzo de la temporada de 1970, tuvo un promedio de tan sólo 0,178. Sustituyendo este valor en la fórmula, hallamos que su habilidad estimada es bastante mayor: el estimador de James-Stein para Munson es 0,247.

De los dos conjuntos de valores, los valores y y los valores z, ¿cuáles expresan mejor las habilidades de los 18 jugadores del ejemplo? Para poder responder con precisión a esta pregunta, tendríamos que conocer la "verdadera habilidad" de cada uno de ellos. Designaremos al verdadero promedio por Θ. Tal valor es incognoscible, y es una abstracción que representa la probabilidad de que un jugador obtenga un tanto en uno cualquiera de sus turnos de bate. A pesar de que Θ es incognoscible, tenemos una buena aproximación de su valor: los resultados obtenidos posteriormente por los bateadores. Bastará considerar el resto de la temporada de 1970, que aportó unas nueve veces más datos que los utilizados para el cálculo de los promedios iniciales. El error estadístico esperado con una muestra tal es lo suficientemente pequeño como para poder despreciarlo, y actuar como si el promedio de temporada fuese la "verdadera habilidad", Θ, del jugador. Este ha sido uno de los motivos de haber elegido promedios de béisbol para el ejemplo. En la mayor parte de los problemas, el verdadero valor de Θ no puede determinarse.

Un criterio de evaluación de ambos grupos de estimaciones puede consistir en contar sus respectivos éxitos y fracasos. Para 16 de los 18 jugadores, el estimador z de James-Stein es más próximo que el promedio observado y al valor "verdadero", o promedio de temporada, Θ. Un método más cuantitativo de comparar ambas técnicas consiste en hallar el error cuadrático total de cada estimador; para ello se determina primero el error real cometido en cada predicción, dado por $(\Theta - y)$ y $(\Theta - z)$, para cada jugador. A continuación, cada uno de estos valores se eleva al cuadrado, y se suman los valores cuadráticos obtenidos. Los promedios observados y tienen un error cuadrático total de 0,077, mientras que el error cuadrático total de los estimadores James-Stein es solamente de 0,022. Según este criterio de comparación, el método de Stein es 3,5 veces más preciso. Puede demostrarse que, para los datos de partida, el valor 3,5 está próximo a la razón esperada de los errores cuadráticos totales de los dos métodos. Así que no es que hayamos tenido pura suerte.

Supongamos que un estadístico hace un muestreo aleatorio de los automóviles de Chicago, y que descubre que, de los primeros 45 observados, nueve son de importación, y los restantes 36, de fabricación nacional americana. Se desea estimar la verdadera proporción de coches importados en Chicago, valor representado por Θ, y también imposible de observar. El valor observado, 9/45=0,200, constituye una estimación de Θ. Puede obtenerse otra estimación englobando este problema con el de los 18 jugadores de béisbol. Introduciendo el valor 0,200 en la ecuación usada en ese problema, se obtiene para la razón de coches importados el valor de 0,251. (En realidad, la introducción de un nuevo valor modifica el gran promedio $\bar{y}$ y altera ligeramente el coeficiente de constricción c. Los cambios son, sin embargo, pequeños; el valor corregido de z es 0,249.)

En este caso, se tiene la fuerte impresión, intuitiva, de que la mejor predicción la dará el promedio observado, y no el estimador de James-Stein. Más aún, todo este segundo método parece absurdo. Pues, ¿qué relación hay entre los promedios del béisbol y la importación de automóviles? En casos así es donde mejor se aprecia la paradójica naturaleza del teorema de Stein, y donde resulta más incómoda. El teorema es tan válido para los 19 problemas como lo era para los primeros 18. Nada en su enunciado



LA DISTRIBUCION NORMAL de una variable aleatoria en torno a su valor medio fundamenta el método de estimación mediante promedios. La distribución normal queda definida por dos parámetros: la media, Θ, que sitúa el pico central de la distribución, y la desviación típica, σ, que mide el grado de dispersión de los datos. Al definir la distribución se supone que la variable x puede tomar cualquier valor del eje x. El valor más verosímil de x es, por definición, la media Θ. La probabilidad de que un valor de x se encuentre en un intervalo dado, como el de extremos a y b, es igual al área limitada entre la curva acampanada y la porción de eje comprendida entre estos puntos.

exige que exista una relación razonable entre los problemas componentes.

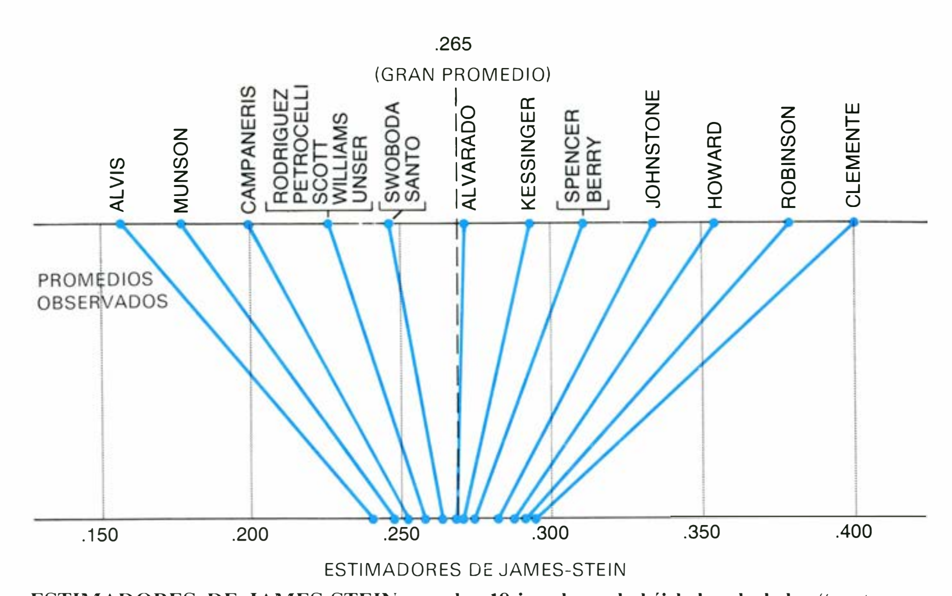
Esta desconcertante indiferencia hacia el sentido común puede ponerse de relieve de otra forma. ¿Qué tiene que ver el promedio observado de Clemente, 0,400, con Max Alvis, el peor de los 18 jugadores considerados? Si Alvis hubiese tenido una buena racha a principios de la temporada, marcando, por ejemplo, 0,444 en lugar del valor real de 0,156, el estimador de James-Stein de Clemente hubiese dado 0,325 en lugar de 0,294. ¿Por qué han de tener los éxitos o fracasos de Alvis ninguna influencia en nuestra estimación de la habilidad de Clemente? (Ni siquiera competían en la misma liga.)

Este es el tipo de objeciones presentadas por los críticos al sistema de Stein. Para poder replicarles se necesita estudiar el método con algo más de cuidado. El cálculo de promedios es un proceso fácil y muy conocido que, al parecer, no precisa justificación. En realidad, no es tan evidente por qué los promedios son a menudo tan útiles para estimar el verdadero centro de gravedad de un proceso aleatorio. La explicación se encuentra en la distribución que tienden a adoptar los valores de la variable aleatoria.

La distribución más generalizada en el trabajo científico es la distribución "normal", descrita por una curva campaniforme; fue estudiada profundamente por Gauss, y se llama en ocasiones distribución gaussiana. Para construirla, se supone que la variable aleatoria puede tomar cualquier valor a lo largo de un cierto eje; después, se hace que la probabilidad de que la variable se encuentre en un intervalo dado sea igual al área limitada entre la curva campaniforme y el intervalo en cuestión. La curva queda completamente determinada por dos parámetros: la media, Θ, que corresponde a la abscisa de la cumbre de la curva, y la desviación típica, que expresa el grado de concentración de los valores de la variable en torno a la media. Es costumbre representar la desviación típica por la letra griega σ. Cuanto mayor sea la desviación típica, tanto más dispersos se encuentran los datos.

En la teoría de la probabilidad, a partir de una media y una desviación típica conocidas se efectúan predicciones. En estadística, el problema discurre a la inversa: a partir de datos observados, el estadístico debe inferir la media Θ y la desviación típica σ.

Supongamos, por ejemplo, que la medición de cierta variable aleatoria x ha dado los cinco valores sucesivos 10,0, 9,4, 10,3, 8,6 y 9,7. Supongamos, además, que se sepa que dichos valores pertenecen a una distribución normal de



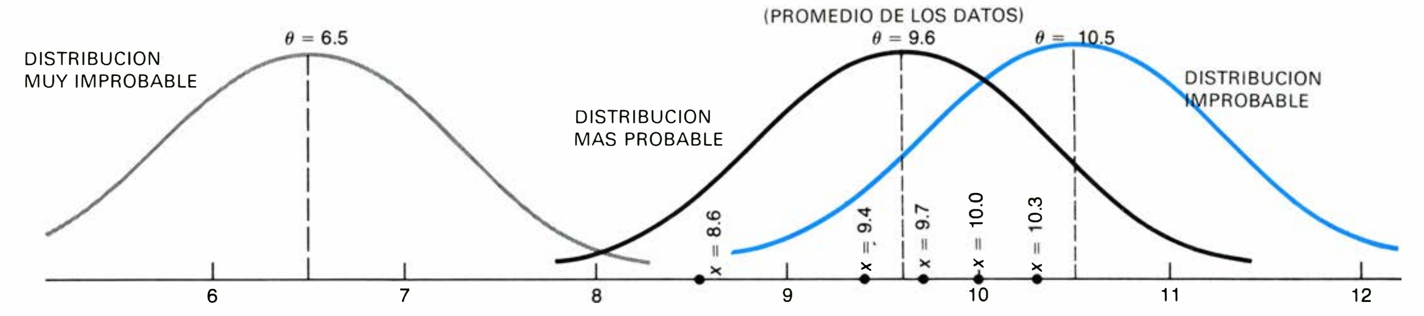
ESTIMADORES DE JAMES-STEIN para los 18 jugadores de béisbol, calculados "contrayendo" sus promedios individuales hacia el "promedio de promedios" global. En este ejemplo, el gran promedio es 0,265, contrayéndose cada promedio individual en aproximadamente el 80 por ciento de su distancia a este valor. Por lo tanto, el teorema que fundamenta el método de Stein afirma que las verdaderas habilidades de los bateadores están más densamente agrupadas de lo que parecen sugerir sus promedios de habilidad preliminares. La ilustración ha sido realizada por Gabor Kiss.

desviación típica igual a 1. ¿Cuál es el valor de la media verdadera, Θ, de la distribución? En principio, la media podría tener cualquier valor, pero ciertos valores son más verosímiles que otros. Por ejemplo, una media de 6,5 exigiría que todos los valores observados se encontraran en una cola de la curva, y ninguno cerca de su centro. Gauss demostró que, entre todos los valores posibles para la media, el valor promedio de los datos observados, $\bar{x}$ (que en este caso tiene el valor 9,6), hace máxima la probabilidad de obtener los datos observados en realidad. En este sentido, el promedio es la estimación más verosímil de la media. Gauss construyó la distribución normal de modo que tuviera esta propiedad.

Existe una razón más, también señalada por Gauss, para tomar el valor promedio como estimación óptima de la media, Θ, inobservable. Gauss hizo notar que el promedio de los datos es un estimador no "sesgado" de la media, en el sentido de que no favorece ningún valor especial de Θ. Con más precisión, el promedio no está sesgado porque el valor esperado de $\bar{x}$ es igual al verdadero valor de Θ, cualquiera que sea éste. Existen infinitos estimadores insesgados de Θ, ninguno de los cuales da una estimación perfecta. Gauss demostró que entre todas las funciones lineales de las observaciones que proporcionan estimaciones no sesgadas de Θ, el promedio, $\bar{x}$, es la que exhibe el error cuadrático de estimación mínimo. Hacia 1940 se demostró que ninguna otra función insesgada de los datos, sea lineal o no, puede estimar la media Θ más exactamente que el promedio, en lo que a error cuadrático se refiere. A la demostración de tal hecho contribuyó, de manera decisiva, R. A. Fisher, quien demostró hacia 1920 que toda la información referente a Θ que pueda extraerse de los datos se encuentra contenida en el promedio, $\bar{x}$.

En la década de 1930, Jerzy Neyman, Egon S. Pearson y Abraham Wald abordaron el problema de establecer la inferencia estadística sobre bases más rigurosas matemáticamente; su trabajo forma parte de la llamada teoría estadística de la decisión. Eliminaron la hipótesis de estimación insesgada, y examinaron todas las funciones de los datos que pudieran servir como estimadores de la media desconocida, Θ. Estos estimadores se comparan mediante una función de riesgo, que se define como el valor esperado del error cuadrático para cada valor posible de Θ.

Supongamos que se están examinando tres posibles estimadores: el promedio de los datos, $\bar{x}$; la mitad de dicho promedio, $\bar{x}/2$, y la mediana o valor central. Para el promedio y para la mediana, la función de riesgo es constante, lo que, dicho de otro modo, significa que el error cuadrático esperado de la predicción de la media Θ es el mismo, independientemente del verdadero valor de Θ. De las dos funciones de riesgo constantes, la del promedio es uniformemente menor en un factor de aproximadamente dos tercios; evidentemente, el estimador "promedio" es el preferido. En lenguaje de teoría de la decisión, la mediana es un estimador "inadmisible" de Θ, pues existe otro estimador de menor riesgo (error cuadrático esperado) independientemente del valor de Θ. (Es preciso decir, sin embargo, que cuando la distribución de



EN ESTADISTICA, EL PROBLEMA consiste en deducir, a partir de un conjunto de datos, la media verdadera y la desviación típica de su distribución. Aun sabiendo que la distribución es normal y que su desviación típica es 1, la media puede tener, en principio, cualquier valor. Sin embargo, algunos valores son más verosímiles que otros. Por ejemplo, los cinco puntos "dato" (x) mostrados en la figura corresponderían a una distribución normal de media 6,5, sólo si se encontraran a más de dos desviaciones típicas por encima de ella. Puede demostrarse que lo más verosímil es que estos datos hayan sido generados por una distribución de media igual al promedio de los datos, que se denota $\bar{x}$.

los datos no sea la normal, puede ocurrir que se invierta el orden de preferencia.)

Para el estimador $\bar{x}/2$, sesgado hacia el valor $\Theta=0$, la función de riesgo no es constante; este estimador es preciso cuando Θ se encuentre próxima a 0, pero el error cuadrático esperado crece rápidamente cuando la media verdadera se aleja del origen. La fundación de riesgo describe una parábola con mínimo en el punto $\Theta=0$. Si la media es efectivamente cero, la función de riesgo de $\bar{x}/2$ es cuatro veces menor que la del propio promedio. No obstante, para valores grandes de la media, el promedio $\bar{x}$ recupera su superioridad. Usando otros estimadores puede rebajarse en cualquier punto dado la función de riesgo hasta valores inferiores a los del promedio, pero siempre a costa de mayores riesgos en otros lugares.

Resta todavía la posibilidad de que algún otro estimador tenga riesgo uniformemente menor que el promedio. En 1950, Colin R. Blyth, Erich L. Lehmann y Joseph L. Hodges, Jr., demostraron la imposibilidad de la existencia de tal estimador. Dicho de otra forma, el promedio $\bar{x}$ es admisible, al menos al utilizarlo para determinar una media desconocida a partir de un conjunto finito de observaciones.

El teorema de Stein se refiere a la estimación de varias medias desconocidas. No se necesita presuponer relación alguna entre ellas; pueden ser habilidades de un bateador o proporciones de automóviles de importación. Por otra parte, se supone que dichas medias son independientes unas de otras. Al evaluar los méritos de los estimadores para estas medias conviene, de nuevo, emplear una función de riesgo definida como la suma de los valores de error cuadrático esperados de todas las medias individuales consideradas.

La elección primaria y obvia de un estimador de cada una de varias medias es tomar el promedio de los datos referentes a cada una de ellas. Todo el desarrollo histórico de la teoría estadística, comenzando por Gauss y pasando por la teoría de la decisión, viene a decir que el promedio es un estimador admisible en tanto que se trate de estimar una sola media, Θ. Stein demostró en 1955 que el promedio también es admisible para la estimación de dos medias. La paradoja de Stein consiste, simplemente, en que demostró que cuando han de estimarse más de dos medias, hacerlo a partir de sus correspondientes promedios es un procedimiento inadmisible. Sean cuales fueren los verdaderos valores de las medias, existen procedimientos de estimación con riesgo total inferior.

En 1955 Stein pudo tan sólo demostrar su proposición para casos en los que el número de medias, que denominaremos k, es muy grande. En el artículo de 1961, Stein, en colaboración con James, pudo extender el resultado para todo valor de k mayor que 2; además, lo hizo de manera constructiva. Stein y James no sólo demostraron que han de existir estimadores superiores en todo punto a los promedios, sino que además pudieron dar un ejemplo de tal tipo de estimador.

El estimador de James-Stein ya ha sido definido en nuestra exposición sobre la habilidad con el bate. Viene dado por la fórmula $z=\bar{y}+c\,(y-\bar{y})$, siendo y el promedio de un solo sistema de datos, $\bar{y}$ el gran promedio de los promedios, y c el coeficiente de constricción. Existen otras varias expresiones del estimador de James-Stein, pero difieren tan sólo en cuestiones de detalle. Todas tienen en común el coeficiente de constricción c, que es la característica fundamental del estimador de James-Stein.

En el problema del béisbol se manejó el coeficiente c como si se tratase de una constante. En realidad, está determinado por los promedios observados, y por consiguiente no es constante. El coeficiente de constricción viene dado por la fórmula

$$c=1-\frac{(k-3)\sigma^2}{\Sigma(y-\bar{y})^2}$$

En ella, k vuelve a ser el número de medias desconocidas a estimar, σ^2 es el cuadrado de la desviación típica, y $\Sigma(y-\bar{y})^2$ es la suma de los cuadrados de las desviaciones de los promedios individuales y respecto al gran promedio $\bar{y}$.

Examinaremos brevemente el significado de esta fórmula de aspecto poco accesible. Una vez fijados k y σ^2, vemos que el coeficiente de constricción disminuye (con lo que las predicciones de las medias se verán tanto más afectadas) al disminuir la expresión $\Sigma(y-\bar{y})^2$. Por otra parte, c aumenta, tendiendo hacia la unidad, cuando la expresión $\Sigma(y-\bar{y})^2$ aumenta, haciéndose menos drástica la contracción de las estimaciones.

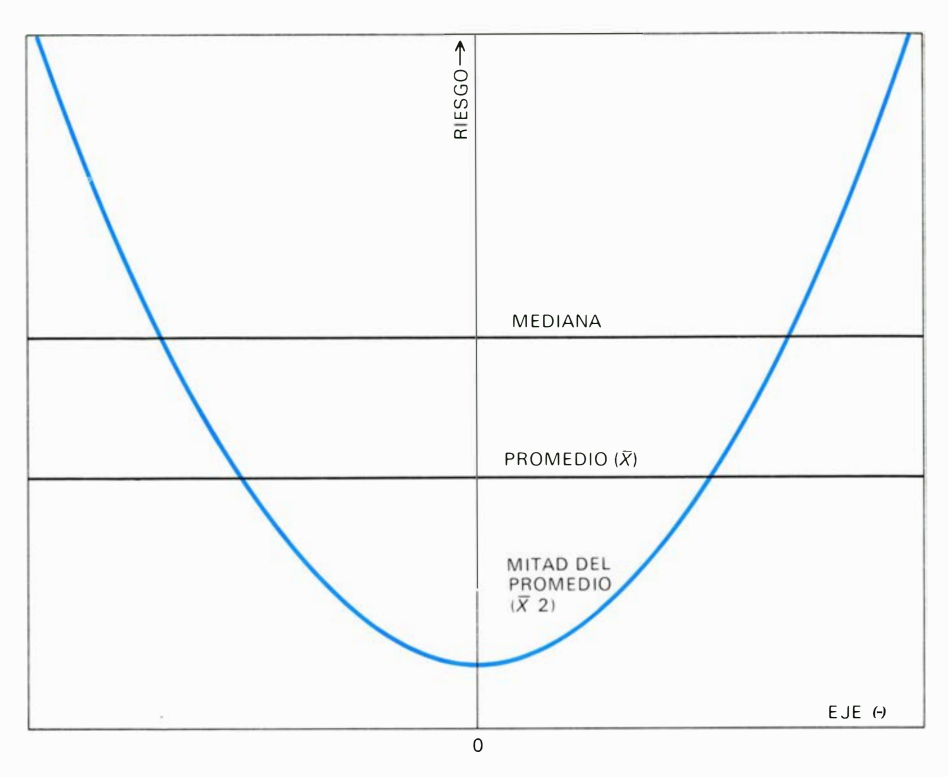
¿Qué revelan estas fórmulas sobre el comportamiento del estimador? De hecho, el procedimiento James-Stein comienza conjeturando que las medias inobservables toman valores cercanos al gran promedio $\bar{y}$. Si los datos observados confirman esta conjetura, es decir, si los promedios de las observaciones no se desvían demasiado de $\bar{y}$, las estimaciones se contraen aún más hacia el gran promedio. Si la conjetura se ve infirmada, no se realiza apenas contracción. Estos ajustes del coeficiente de constricción se efectúan en virtud de la influencia que la distribución de los promedios en torno al gran promedio tiene en la fórmula que determina el valor de c. También influye en el coeficiente de constricción el número de medias a estimar, en virtud del factor $(k-3)$ que aparece en esta misma fórmula. Cuando hay muchas medias, la contracción impuesta por la fórmula es más drástica, pues en tal caso es menos verosímil que las variaciones observadas representen puras fluctuaciones aleatorias.

Calculado así el valor de c, la función de riesgo del estimador de James-Stein es inferior a la de los promedios de las

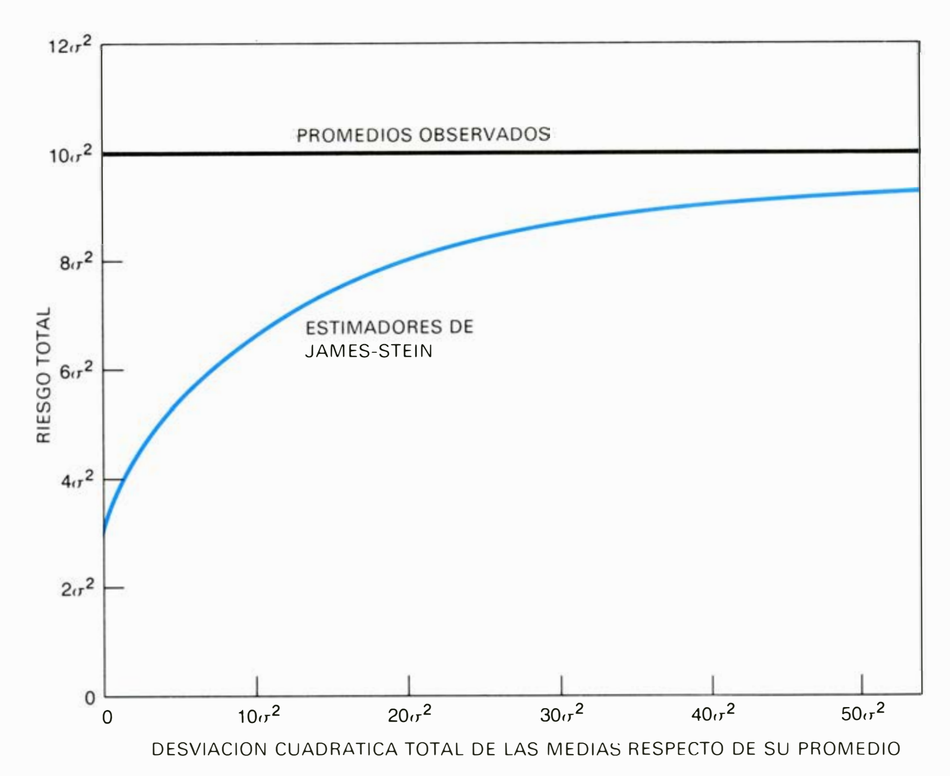
muestras, independientemente de los verdaderos valores Θ de las medias. El riesgo puede disminuir de modo apreciable, especialmente si el número de medias excede de cinco o seis. A diferencia de lo que ocurre con los promedios observados, la función de riesgo no es constante para todo valor de Θ. El riesgo del estimador de James-Stein es mínimo cuando todas las medias verdaderas son iguales. Conforme las medias van alejándose unas de otras, va aumentando el riesgo del estimador, tendiendo hacia el de los promedios observados, aunque sin alcanzarlo jamás del todo. El estimador de James-Stein tan sólo se muestra verdaderamente superior a los promedios cuando las medias verdaderas están próximas entre sí, con lo que la conjetura inicial resulta confirmada. Lo que de todos modos resulta sorprendente es que el estimador se muestre siempre superior a los promedios, aunque sea marginalmente, con independencia de los verdaderos valores de las medias.

En la versión del estimador James-Stein aquí expuesta, se comparan con el gran promedio $\bar{y}$ todos los promedios observados. Este proceder no es el único posible; en otras expresiones del estimador se prescinde por completo de $\bar{y}$. Pero no es posible prescindir de cierta hipótesis inicial, arbitraria en mayor o menor grado, que sirva como origen o punto de referencia del estimador. Como es evidente, los promedios observados no dependen de la elección del origen. Antes de que Stein descubriese su método se tenía la opinión de que los estimadores "invariantes", como los anteriores, habrían de ser preferibles a los que dan predicciones dependientes del origen elegido. La teoría de invariancia, de la que Stein había sido uno de los principales defensores, se vio conmovida en sus cimientos por el contraejemplo de James-Stein. Desde el punto de vista matemático, éste es sin duda el aspecto más perturbador del teorema de Stein. De hecho, la paradoja no se descubrió antes por el prejuicio fuertemente arraigado de que, al estar planteado el problema de estimación sin referencia a ningún origen, debería resolverse de manera análoga, es decir, sin referencia a ningún origen.

En las aplicaciones del teorema de Stein suelen intervenir grandes conjuntos de datos con muchos parámetros desconocidos. Algunas de las dificultades de tales problemas, así como la potencia práctica del método, pueden ilustrarse mediante un ejemplo: el análisis de la distribución de la enfermedad llamada to-



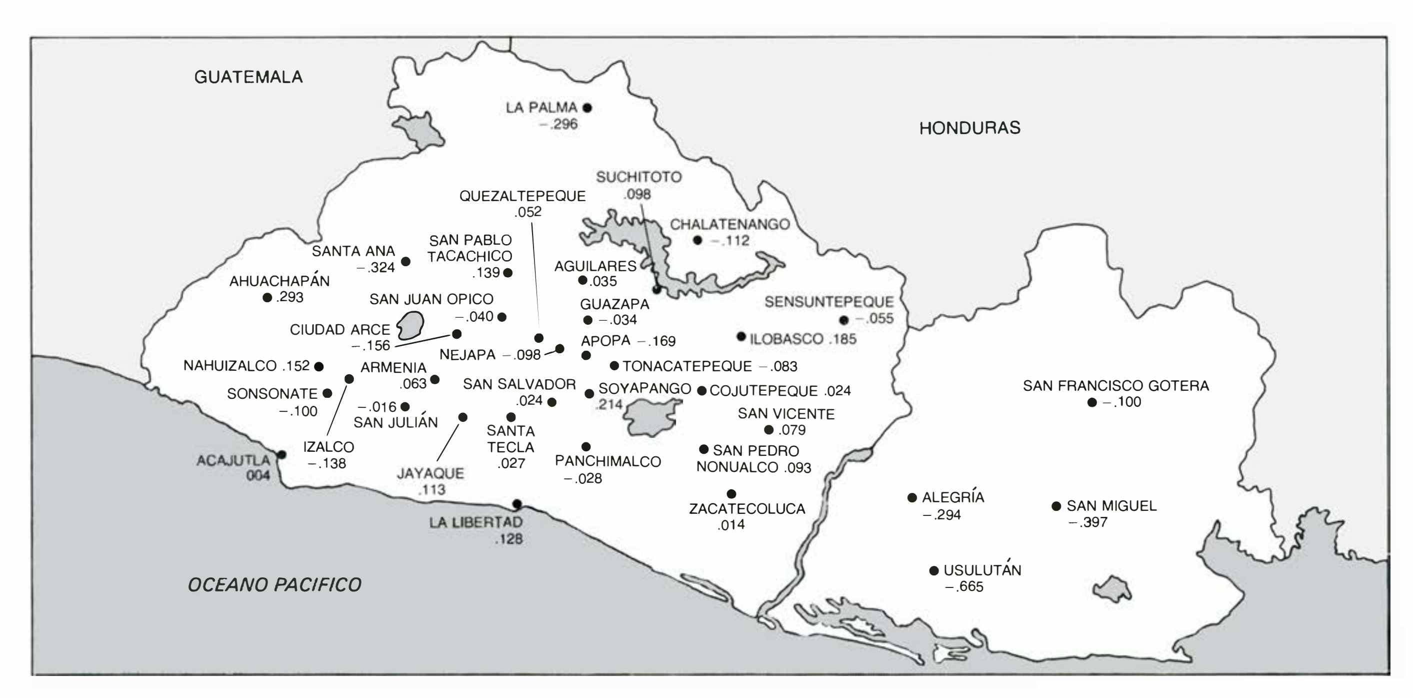
DIFERENTES ESTIMADORES de una única media verdadera, Θ, pueden evaluarse mediante una función de riesgo. El riesgo se define como valor esperado del error cuadrático de la estimación, considerado como función de la media Θ. El promedio de los datos, $\bar{x}$, es un estimador cuya función de riesgo es constante: independientemente del verdadero valor de la media, el valor esperado del error cuadrático es el mismo. La mediana, o valor central, de los datos también tiene riesgo constante, pero su riesgo es mayor en todo punto (por un factor de 1,57) al del promedio. La mitad del promedio ($\bar{x}/2$) es un estimador cuyo riesgo depende del valor real de la media; el riesgo es mínimo si la media se encuentra próxima a cero y aumenta rápidamente al alejarse del origen.



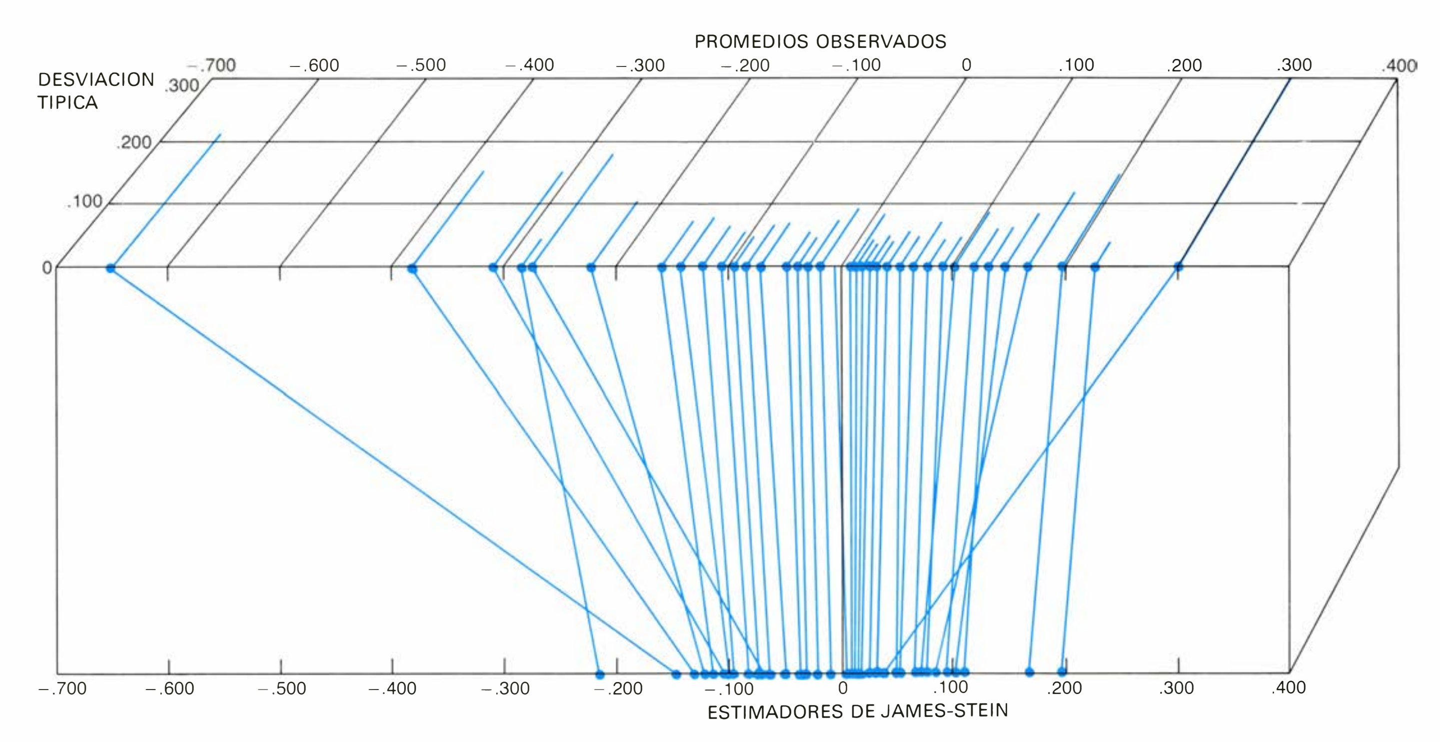
LA FUNCION DE RIESGO TOTAL del estimador de James-Stein es en todo punto menor que la de los promedios individuales observados, tan pronto como el número de medias a estimar sea mayor que dos. En este ejemplo hay 10 medias desconocidas. El riesgo es mínimo cuando todas las medias se apiñan en un solo punto. Conforme las medias se separan unas de otras el riesgo de los estimadores de James-Stein aumenta, tendiendo hacia el de los promedios, aunque sin alcanzarlo.

xoplasmosis en la república salvadoreña.

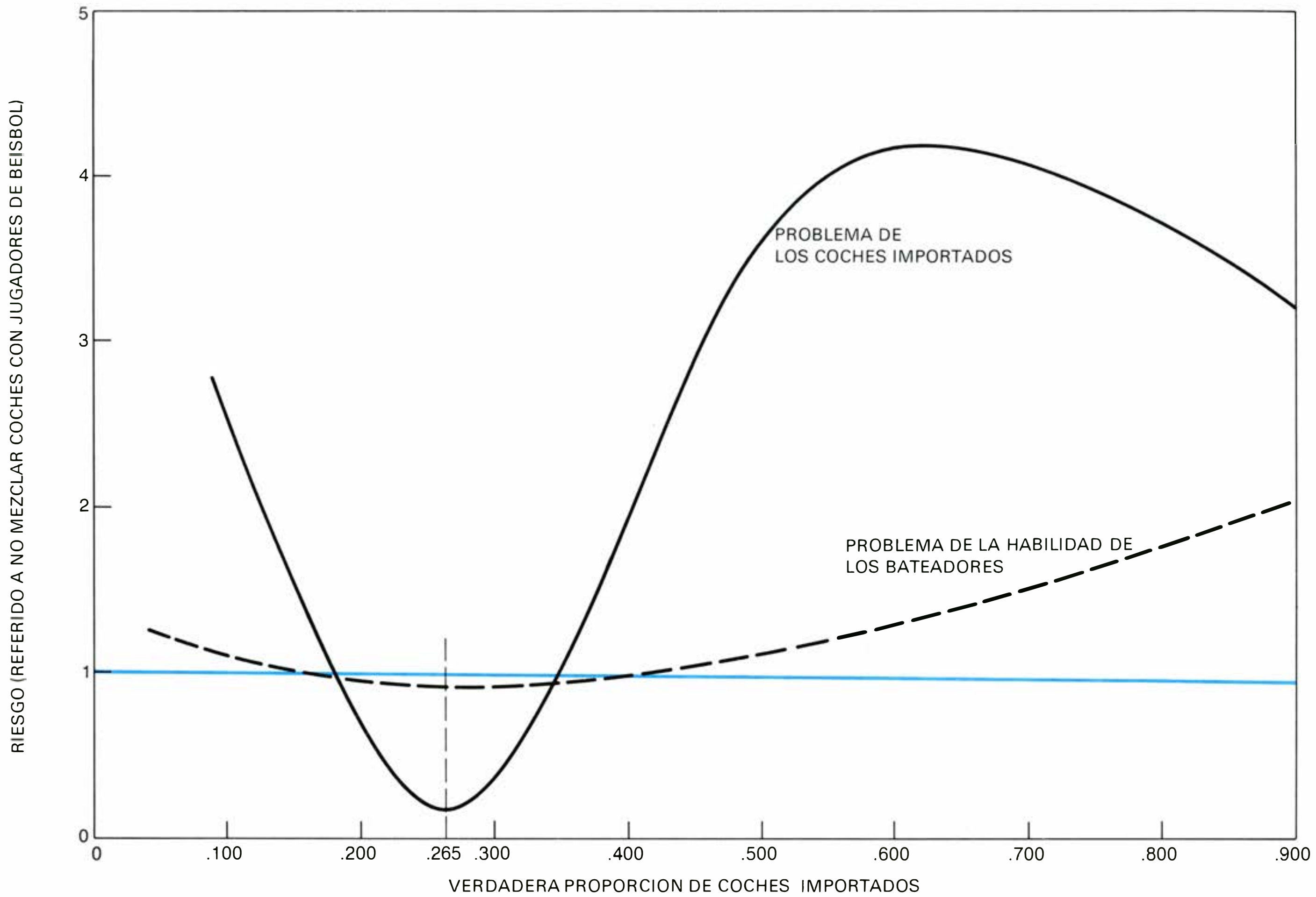
La toxoplasmosis es una enfermedad de la sangre, endémica en muchos países de América Central y, en general, en regiones tropicales. En El Salvador se examinó a aproximadamente 5000 personas, procedentes, en número variable, de 36 ciudades. Conviene expresar la morbosidad por ciudad comparándola con el valor nacional (es decir, con el gran promedio $\bar{y}$). Así, por ejemplo, una indicación de 0,050 representa una ciudad en la que la incidencia de la enfermedad es un 5 por ciento superior al promedio nacional. Las tasas medidas tienen una distribución aproximadamente normal. Las desviaciones típicas de dichas distribuciones son conocidas, pero varían de ciudad a ciudad, dependiendo inversamente del tamaño de la muestra examinada en cada una. La tarea del estadístico con-



MORBOSIDAD DE TOXOPLASMOSIS, una enfermedad de la sangre, medida en 36 ciudades de la República de El Salvador. La incidencia observada en cada ciudad puede tomarse como estimación de la verdadera incidencia, que es inobservable. Esta morbosidad observada tiene distribución normal, y su desviación típica está determinada por el número de personas examinadas en esa ciudad. Las mediciones se expresan en forma de diferencias respecto a la morbosidad nacional (que es el promedio de las morbosidades observadas en todas las ciudades). Así pues, una ciudad cuya incidencia observada sea –0,040 tendrá una morbosidad un cuatro por ciento menor que la global de todo el país.



CONTRACCION de los porcentajes observados de toxoplasmosis de acuerdo con la técnica de estimación de James-Stein. La distribución aparente de la enfermedad queda muy alterada. El coeficiente de contracción no es el mismo para todas las ciudades, sino que depende de la desviación típica del porcentaje medido en cada ciudad. Una desviación típica grande significa que la medida está basada en una muestra pequeña, y, por tanto, sometida a fluctuaciones aleatorias grandes; dicha medida se comprime más que las otras. En los datos referentes a El Salvador, las observaciones extremas tienden a estar en correlación con las desviaciones típicas máximas, lo que hace temer su escasa fiabilidad. Puede demostrarse que los estimadores de James-Stein implican, en comparación con los promedios observados, un menor error total.



PROBLEMAS SIN RELACION MUTUA pueden mezclarse en el método de análisis de Stein, pero a riesgo de aumentar el error. Por ejemplo, podría añadirse al conjunto de 18 promedios de béisbol un décimonono número que indicase la proporción de coches de importación observada en Chicago. Se podría calcular entonces nuevos estimadores de James-Stein para el sistema mixto de jugadores y automóviles, basados en el gran promedio de estos 19 números. En el enunciado del teorema de Stein nada impide proceder así, aunque su evidente arbitrariedad ha sido justificablemente criticada. En realidad, la inclusión de datos sin relación sólo podrá reducir la función de riesgo si la proporción de coches importados resulta hallarse cercana al promedio, 0,265, de habilidades de los bateadores; en caso contrario, el error de estimación esperado aumento tanto para los coches de importación en Chicago como en el caso de que se trate de los dieciocho jugadores.

siste en estimar la media verdadera Θ de la distribución de cada ciudad, a partir de la incidencia observada, y.

En este caso, la forma adecuada del estimador James-Stein es $z=cy$. La simplificación, que hemos introducido nosotros, se posibilita por la forma en que se ha elegido expresar las observaciones y. Se las define de modo que el gran promedio $\bar{y}$ sea cero, y todos los términos en que intervenga el gran promedio $\bar{y}$ desaparezcan, por tanto, de la fórmula. Por otra parte, el proceso de estimación es ahora más complicado, al ser el coeficiente de constricción distinto para cada ciudad, variando en proporción inversa a la desviación típica de y de cada ciudad. La dependencia del coeficiente de constricción respecto de la desviación típica tiene una justificación sencilla e intuitiva. Una desviación típica grande implica un alto grado de azar e incertidumbre en las correspondientes observaciones. Si la morbosidad observada es insólitamente grande, quizá sea más razonable atribuirla a fluctuaciones aleatorias dentro de la distribución normal que a un valor genuinamente grande de la media verdadera Θ. En tal caso se justifica aplicar un coeficiente de constricción pequeño, que reduzca drásticamente el valor anormal.

Para reforzar aún más estas consideraciones, volvamos por un instante al béisbol. En la temporada de 1893, Frank O'Connor fue "pitcher" del equipo de Filadelfia. En su carrera como jugador de primera división hubo de tomar el bate en dos ocasiones, ambas con éxito. Su promedio observado es, por tanto, 1,000. La regla de James-Stein para los 18 jugadores del ejemplo inicial estima que la verdadera habilidad de O'Connor es igual a $0{,}265+0{,}212\ (1{,}000-0{,}265) = 0{,}421$ (despreciando el efecto de los nuevos datos sobre el coeficiente de constricción). Esta estimación es absurda, pero no lo es tanto como 1,000. Alcanzar un promedio perfecto en dos ensayos no está en contradicción, en modo alguno, con un verdadero valor comprendido entre 0,242 y 0,294, que es el estimado para los otros jugadores. El coeficiente c aplicado a O'Connor debería ser todavía menor, para compensar el menor número de datos referentes a él.

En el caso de las observaciones de El Salvador, la mayoría de coeficientes de constricción son muy suaves; fluctúan en general entre 0,6 y 0,9, aunque unos cuantos oscilan entre 0,1 y 0,3. ¿Qué conjunto de cifras deberíamos elegir: las estimaciones de James-Stein, o las morbosidades observadas? La respuesta depende mucho de la aplicación que vaya a dárseles.

Si el Ministerio de Sanidad de El Salvador se propusiese construir una red de hospitales locales para enfermos de toxoplasmosis, las estimaciones de James-Stein ofrecerían probablemente una orientación más segura, pues el valor esperado del error cuadrático total es, en este ejemplo, unas tres veces menor para estas estimaciones. Es importante tener presente que el error previsto se ha calculado sumando los de todas las ciudades. Un hospital determinado podría estar mal emplazado o no tener la capacidad apropiada, pero la suma de todos los desajustes sería menor para los estimadores

de James-Stein que para las morbosidades observadas.

También es verosímil que resulten preferidas las estimaciones del método James-Stein a la hora de determinar la ordenación de las medias verdaderas. Destaca, a este respecto, que la ciudad de máxima morbosidad (de acuerdo con los promedios medidos, y) ocupe el duodécimo lugar en el método de James-Stein. La drástica reducción de la estimación se debió a que en esa ciudad la muestra fue muy reducida. Esta información podría ser útil si solamente se dispusiera de fondos para un hospital.

Supongamos que un epidemiólogo desease investigar la correlación entre la morbosidad verdadera de cada ciudad y factores tales como la pluviosidad, la temperatura, la densidad de población, la altitud, etc. Volverían a ser preferibles los estimadores de James-Stein. Un cálculo somero mostraría que darían mejores aproximaciones en alrededor del 70 por ciento de los casos.

Hay una finalidad para la cual la morbosidad observada puede muy bien ser superior al estimador de James-Stein: cuando se estudia aisladamente una sola ciudad. Como ya hemos visto, el método de James-Stein da mejores resultados cuando hay muchas ciudades, y hace menor el error total de estimación de la suma correspondiente a todas las ciudades. No puede demostrarse, sin embargo, que el método de Stein sea superior al aplicarlo a una ciudad concreta; en realidad, la estimación de James-Stein puede ser francamente peor.

Pueden cometerse graves errores al estimar por el método de Stein la media verdadera de una ciudad aislada cuando dicha media tenga valor atípico. La idea subyacente en el método es reducir el riesgo global suponiendo que las medias verdaderas son más parecidas entre sí que los datos observados. Esta suposición puede degradar las estimaciones de medias genuinamente atípicas. Ahora podemos comprender por qué no deben mezclarse los automóviles importados en los mismos cálculos que los 18 jugadores de béisbol: hay una gran probabilidad de que los automóviles formen una población atípica.

Supongamos que se decide correr el riesgo y agrupar en uno solo los 19 problemas. Podemos calcular entonces el error cuadrático total esperado en función del verdadero porcentaje de coches importados. Resulta que el riesgo, tanto para los jugadores como para los autos, se reduce solamente si ocurre que el porcentaje de importación cae dentro de la gama de valores de habilidad de los bateadores: de lo contrario, el riesgo de error aumenta para ambos problemas.

Decidir si una media concreta es o no "típica" es una cuestión sutil, cuyas implicaciones aún no se comprenden del todo. Volvamos al problema de la toxoplasmosis en El Salvador, y fijémonos en la ciudad de Alegría, que muestra la quinta cifra más baja de incidencia observada de la enfermedad: $-0{,}294$. Se trata de una de las cuatro ciudades situadas al este del río Lempa incluidas en la encuesta: todas ellas tienen valores netamente negativos de la morbosidad medida. Es válido suponer que no se trata de coincidencias, y que la proporción de toxoplasmosis al este del río Lempa es verdaderamente inferior. Un estimador de James-Stein que dé información consolidada para todo el país puede ser mucho menos que óptimo para estas ciudades. Nosotros hemos desarrollado técnicas para aprovechar este tipo de información extra, pero la teoría que sustenta dichas técnicas sigue siendo rudimentaria.

Con un poco de malicia, un aficionado al béisbol podría caer en la cuenta de que así como la habilidad individual de cada jugador puede representarse mediante una curva de Gauss, también las verdaderas habilidades de todos los jugadores de primera división tienen una distribución aproximadamente normal. La media de esta distribución es 0,270 y la desviación típica 0,015. Con esta valiosa información adicional, que los estadísticos llaman distribución previa, se puede construir una estimación mejor de la verdadera habilidad de los jugadores. Este nuevo estimador, que denotaremos Z, se define por la ecuación $Z=m+C(y-m)$, siendo y la habilidad observada de cada jugador, aunque aquí el gran promedio $\bar{y}$ se ha sustituido por m, que es la media de la distribución previa, cuyo valor conocido es 0,270. Hay además un coeficiente de constricción distinto, C, que depende de modo sencillo de la desviación típica de la distribución previa (igual a 0,015).

Este procedimiento no es un refinamiento del método de Stein; por el contrario, es 200 años anterior, y es la formulación matemática de un teorema publicado (póstumamente) por el reverendo Thomas Bayes en 1763. Bayes pudo demostrar que este estimador hace mínimo el error cuadrático previsto asociado al carácter aleatorio tanto de los promedios observados (y) como de las medias verdaderas, (Θ).

El parecido de las fórmulas de los estimadores de Bayes y de James-Stein llama poderosamente la atención. De hecho, cuando el número de medias consideradas tiende a infinito, ambas ecuaciones se hacen idénticas. Los dos coeficientes de constricción c y C convergen hacia un mismo valor, y el gran promedio $\bar{y}$ se hace igual a m precisamente cuando se hace intervenir en el cálculo a todos los jugadores. El procedimiento de James-Stein tiene, sin embargo, una importante ventaja sobre el método de Bayes: es aplicable sin necesidad de conocer la distribución previa. En realidad, ni siquiera es preciso suponer que las medias que se quiere estimar tengan una distribución normal. Por otra parte, la ignorancia tiene un precio, que se paga con una precisión inferior de las estimaciones. Nosotros hemos podido demostrar que el método de James-Stein aumenta la función de riesgo en proporción a $3/k$, siendo k, como antes, el número de medias por estimar. El riesgo extra es despreciable cuando k pasa de 15 o 20, y es tolerable para valores tan bajos como $k=9$.

En este contexto histórico, el estimador de James-Stein puede considerarse como una "regla de Bayes empírica", según la expresión acuñada por Herbert E. Robbins, de la Universidad de Columbia. En unos trabajos que comenzaron en 1951, Robbins demostró que puede alcanzarse el mismo riesgo mínimo asociado a la regla de Bayes, aun sin conocer la distribución previa, siempre que el número de medias por estimar fuese muy grande. La teoría de Robbins se reconoció de inmediato como una aportación fundamental: mucho más lenta ha sido la aceptación del resultado de Stein, con el que guarda estrecha relación.

El estimador de James-Stein no es el único del cual se sepa que es mejor que los promedios muestrales. En realidad, el propio estimador de James-Stein es inadmisible. Su defecto consiste en que el coeficiente de constricción c puede tomar valores negativos, y entonces puede ocurrir que las medias, en vez de contraerse hacia el gran promedio, se dispersen. En tales casos, basta cambiar c por cero para obtener una estimación mejor. A su vez, este estimador también es inadmisible, aunque todavía no se conoce un estimador uniformemente superior.

Continúa la búsqueda de nuevos estimadores. Los esfuerzos más recientes se están concentrando en la obtención de resultados análogos a los de Stein para problemas con distribuciones distintas de la normal. Varios métodos de investigación, entre los que se encuentran los de Stein, Robbins y métodos bayesianos más formales, parecen estar convergiendo hacia una potente teoría general de estimación de parámetros.

Juegos matemáticos

La inducción matemática y su semejanza con la caída de una hilera de piezas de dominó

Martin Gardner

"Es cierto", respondió, "no hacen falta ojos para verlo".

PLATON, *La República*, Libro V

No hay forma de estar absolutamente seguro. ¿Saldrá el sol el primer día del mes de enero del año 2000? Siempre es posible que el mundo se acabe antes, ya por decreto divino, ya por alguna catástrofe natural. Quizás un cometa gigante (como el de la mitología de Immanuel Velikovsky) detenga el movimiento de rotación terrestre, y el sol se inmovilice sobre Gibeón. Afirmar que, como cada día, el sol saldrá también el 1 de enero del año 2000, es un excelente pronóstico; pero esto es lo más que puede decirse. Como en el caso anterior, al pasar de un número finito de amaneceres ocurridos a un conjunto infinito futuro, o al menos, a un conjunto futuro de gran número de elementos, se realiza una inducción empírica.

En matemáticas se dispone de una técnica análoga, conocida por el nombre de inducción matemática o inducción completa, que implica también saltar de un conjunto finito de casos conocidos a un conjunto mayor o a un número infinito de casos. A diferencia de la inducción empírica, esta técnica matemática es completamente deductiva: una demostración por inducción completa es todo lo rigurosa que pueda serlo una demostración matemática.

Una demostración por inducción matemática requiere, en primer lugar, una serie de enunciados (por lo común, aunque no necesariamente, infinita) que pueda ponerse en correspondencia biunívoca con la sucesión de los números enteros positivos. Segundo, ha de establecerse que los enunciados estén relacionados entre sí por lo que Bertrand Russell llamó "propiedad hereditaria", a saber: si cierto enunciado es verdadero, entonces su sucesor –el enunciado "siguiente"– también es verdadero. Tercero, hemos de probar que el primer enunciado es verdadero. Resulta entonces con férrea certidumbre que todos los enunciados son verdaderos.

Existe cierto paralelismo entre una demostración por inducción y una hilera de ladrillos o dominós, de pie sobre uno de sus bordes: si cae el primero de la fila, caerán todos. Hugo Steinhaus comparaba la inducción matemática con una pila de sobres, en cada uno de los cuales podía leerse la siguiente nota: "Abra el sobre siguiente, lea la orden que contiene, y ejecútela". Si se está obligado a obedecer la orden del primer sobre, se tendrán que abrir todos los demás y ejecutar todas las órdenes.

Cientos de problemas de matemáticas recreativas se estudian, en su caso general, por inducción completa. ¿En cuántas porciones puede, como máximo, cortarse una tarta mediante n cortes rectos? ¿Cuál es el mínimo número de movimientos para trasladar n discos en el problema de la Torre de Hanoi? En este mes nos ocuparemos de una clase de retorcidos rompecabezas en los que la aplicación al caso general del método de inducción es menos conocida y está plagada de curiosos riesgos.

Comenzaremos por el viejo problema de los tres sombreros. A tres hombres, *A, B* y *C*, que mantienen los ojos cerrados mientras tanto, se les pone un gorro rojo o un gorro negro. Al abrir los ojos, cada hombre ve los gorros de los demás, pero no el suyo; si ve un gorro rojo, tiene la obligación de alzar la mano. Tan pronto como sepa el color de su gorro, ha de decirlo así.

Supongamos que los tres sombreros sean rojos. Los tres hombres alzan la mano. Tras cierto lapso de tiempo, *C*, que es más listo que los demás, dice: "Mi gorro es rojo". ¿Cómo ha podido saberlo?

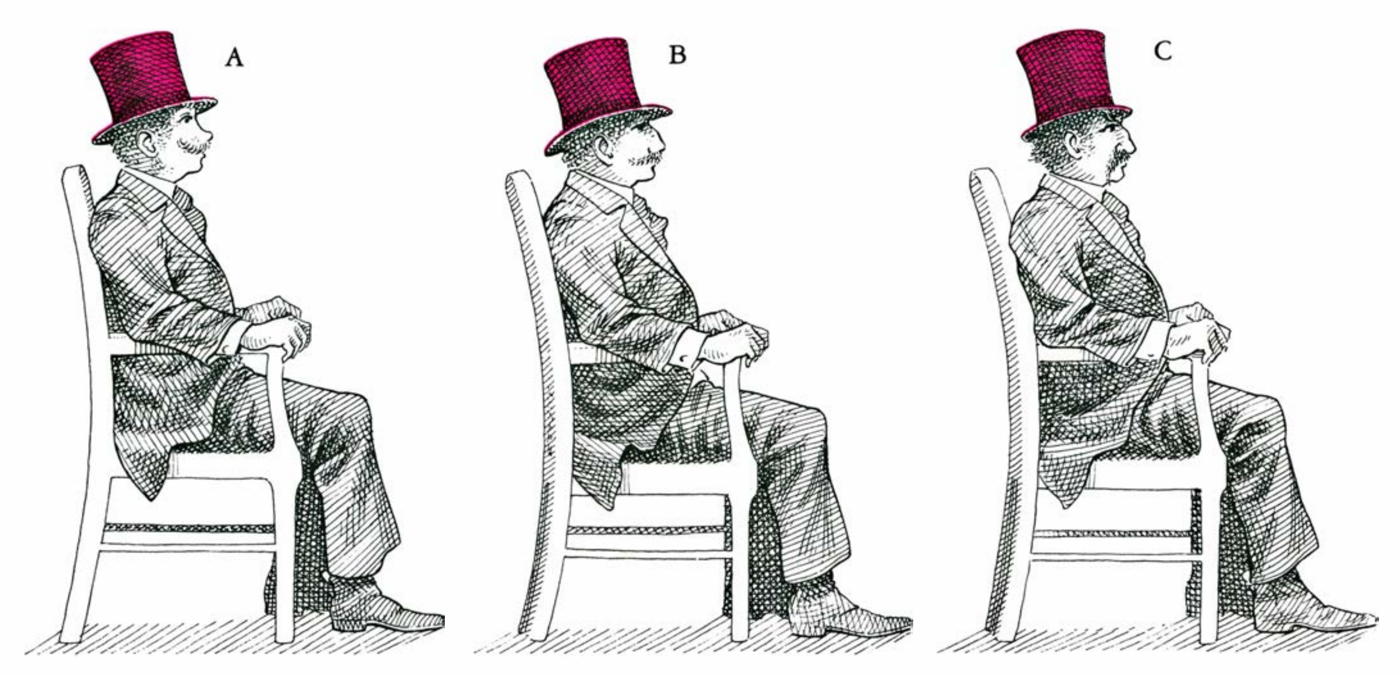
C razona de la manera siguiente: "Supongamos que mi gorro fuese negro. Entonces, *A*, al ver mi gorro negro, sabría en seguida que su gorro es rojo, pues si no, ¿por qué habría *B* de alzar la mano? *B* podría razonar igual; sabría también que su gorro es rojo. Sin embargo, ni *A* ni *B* han dicho nada. Sus dudas solamente tienen fundamento si ellos ven en mí un gorro rojo. Así pues, mi gorro es rojo".

Estudiemos ahora el problema con cuatro hombres, todos los cuales reciben gorros rojos. Si el cuarto, *D*, es más inteligente que los otros, razonará así: "Supongamos que mi gorro sea negro. Los otros tres han alzado la mano porque ven gorros rojos. Este es precisamente el caso precedente. Tras un tiempo razonable, *C*, que es el más listo de los otros tres, habrá descubierto que su gorro es rojo, y así lo dirá". *D* aguarda entonces a que hable *C*. El silencio de *C* le prueba a *D* que su gorro es rojo.

Evidentemente, este proceso puede generalizarse. Si hubiesen cinco hombres, *E* podrá saber que su gorro es rojo, pues si fuese negro, la situación se reduce al caso anterior; tras un lapso de tiempo adecuado, *D* sabría que su gorro es rojo. Puesto que *D* guarda silencio, todos los gorros, y también el suyo, son rojos. Y así sucesivamente con cualquier número de personas. Por inducción matemática nos vemos forzados a concluir que si a la totalidad de n hombres se les colocan sombreros rojos, el más inteligente de todos ellos terminará deduciendo que su gorro es rojo.

Normalmente esta generalización es puesta en tela de juicio. El problema requiere varias nebulosas hipótesis sobre grados de inteligencia y lapsos de tiempo razonables, cada vez mayores, y acaba siendo irreal. Presumiblemente, si fuesen 100 hombres, después de algunas horas el más inteligente sabría que su sombrero es rojo; un poco más tarde, lo sabría el segundo en inteligencia, y así sucesivamente, hasta llegar a los dos más estúpidos.

Pueden eliminarse todas estas vaguedades, y darle al problema una forma más precisa. Hay tres hombres y cinco sombreros. Tres sombreros son rojos, y, dos, negros. Se supone que todos los hombres son honestos y "racionales", en el sentido de que pueden efectuar rápidamente cualquier deducción, por muy complicada que sea. Como antes, los hombres cierran los ojos, y un "árbitro" cubre con un gorro sus cabezas. En lugar de pedírseles que levanten la mano si ven gorro rojo, se les pregunta por or-



El problema de los sombreros

den: "¿Sabe Ud. de qué color es su gorro?"

A dice la verdad contestando "no". También *B* contesta "no". *C* dice: "Sí. Mi gorro es rojo". ¿Cómo lo ha sabido?

Un sorprendente aspecto que ofrece este problema es que *C* puede contestar "sí" ¡aunque sea ciego! Además, no es necesario que *B* vea el gorro de *A*. Imaginemos a los tres hombres sentados en una hilera de asientos, como se muestra en la figura de esta página. Cada uno ve los gorros de los sentados más adelante. *C*, que ocupa la tercera posición, es ciego, en el sentido de que no ve el gorro de nadie.

C razona del modo siguiente: "*A* tan sólo podrá decir "sí" cuando vea dos gorros negros. Al decir que no, resulta claro que los sombreros *B* y *C* no son ambos negros. Supongamos que mi gorro sea negro. *B* puede ver que lo es. Así pues, tan pronto como *B* oiga que *A* dice "no", sabrá que su gorro es rojo. (Pues de lo contrario, los gorros *B* y *C* serían negros, y *A* habría dicho "sí".) La única explicación de que *B* haya dicho "no" es que vea en mí un gorro rojo. Por tanto, puedo contestar "sí".

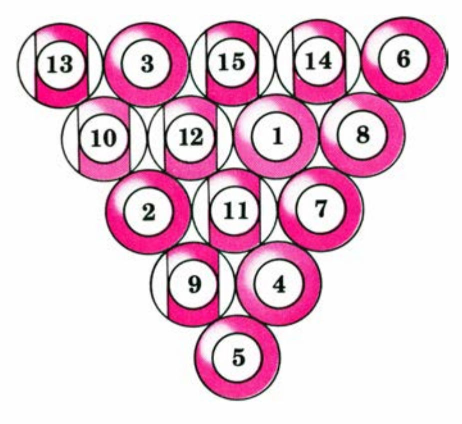
Lo mismo que el anterior, este problema se generaliza sin dificultad a *n* hombres sentados en una hilera de asientos, disponiéndose ahora de *n* sombreros rojos y *n*–1 sombreros negros. Supongamos que un cuarto personaje, *D*, esté sentado delante de *C*. Todos los gorros son rojos. *D* puede razonar que si su gorro fuera negro, al verlo, los sentados a su espalda sabrían que para ellos solamente restan dos, con lo que el problema se reduce al precedente, ya resuelto. Después de que *A* y *B* dijeran "no", *C* contestaría "sí". Pero *C* también contesta "no", lo que le demuestra a *D* que su propio gorro también es rojo. La solución se generaliza sin dificultad, por inducción matemática, a *n* personas. Si todas reciben sombreros rojos, todas contestarían que no, excepto la *n*-ésima, quien sabría que su gorro es rojo.

Podemos ahora plantear una cuestión más difícil. Imaginemos otra vez a los tres hombres sentados en hilera, y supongamos que el árbitro distribuye entre ellos una combinación cualquiera de gorros, extraída del conjunto de cinco. Se interroga a los sujetos en orden de "ceguera" creciente *(A, B, C.)* ¿Podrá siempre uno de ellos contestar que sí? ¿Podría generalizarse la situación a *n* personas y a un conjunto de *n* gorros rojos y *n*-1 gorros negros? ¿Se obtendrá en todos los casos una respuesta afirmativa antes de la *n*-ésima pregunta, o en ella, independientemente del reparto de gorros?

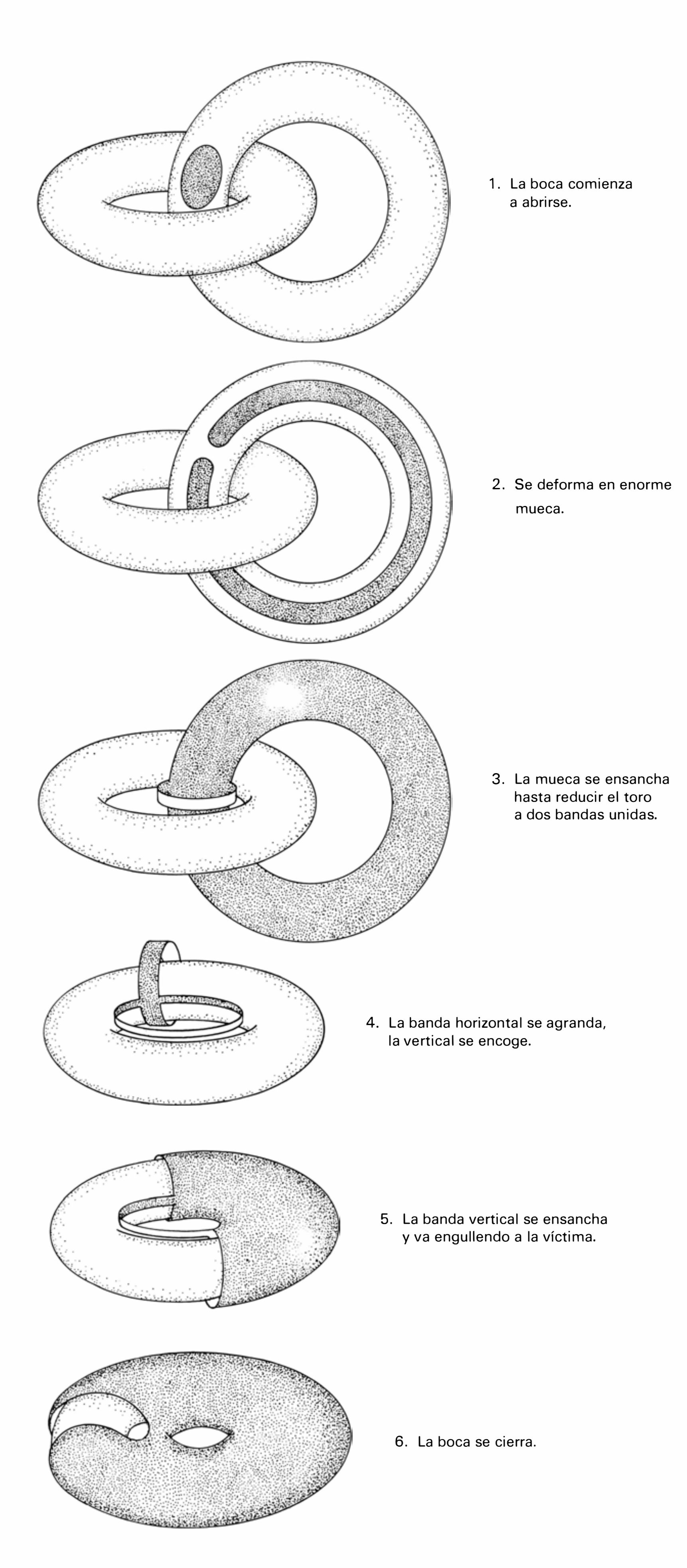
En la mayoría de problemas de este tipo se tropieza con una curiosa paradoja. Consideremos el caso de tres personas, todas con gorro rojo, y donde cada una ve a las otras dos. *A* y *B* responden que no. *C* contesta que sí. ¿Qué necesidad hay de interrogar a *A*? Antes de que se le pregunte, *B* y *C* saben que *A* tiene que contestar "no"; *B* lo sabe porque ve el gorro rojo de *C*, y *C* lo sabe porque ve el gorro rojo de *B*. Puesto que *B* y *C* saben cuál habrá de ser la respuesta de *A*, ¿de qué forma puede aportar nueva información la pregunta que se le formula a *A* y su respuesta? Por otra parte, si el interrogatorio comenzase por *B*, a *C* no le sería posible efectuar su razonamiento. ¿Puede el lector explicar esta paradoja aparente?

Sombreros de dos colores equivalen a sombreros marcados 0 y 1, los dígitos del sistema de numeración binaria. Hay docenas de problemas íntimamente relacionados con los de sombreros, en los que intervienen más de dos colores, y que son más fáciles de comprender si en lugar de colores se usan enteros positivos en base diez. El año pasado, David Gale, matemático de la Universidad de California en Berkeley, me envió el siguiente juego bipersonal.

El árbitro elige un par cualquiera de números enteros positivos consecutivos. Se pegan sendos discos con dichos números en la frente de dos personas. Ambas son honestas y racionales, y cada una ve el número de la otra, pero no el propio. Ambas saben, y saben que la



Solución del problema del billar



Cómo uno de los toros devora al otro

otra sabe, que los números son consecutivos.

El árbitro pregunta por turno a cada jugador si conoce su número; el interrogatorio continúa hasta que uno de ellos diga que sí. No es difícil demostrar, mediante la mágica inducción, que finalmente el jugador de número más alto, n, será el que primero diga que sí, y que dicho sí vendrá en respuesta a la n-ésima o n-1-ésima pregunta. Se invita a los lectores a que analicen el juego y determinen en qué condiciones la respuesta afirmativa se producirá en la n-ésima pregunta y cuándo en la n-1-ésima. Sólo es preciso considerar dos variables, a saber, si el número mayor es par o impar y si se pregunta en primer lugar al jugador de este número o al otro.

La paradoja del juego de los sombreros aparece aquí en forma todavía más llamativa. Me inspiraré ahora en la carta de Gale. Supongamos que los números sean 99 y 100, y que se pregunta en primer lugar al jugador del 100. Este dirá que sí en la centésima pregunta. Pero, ¿qué necesidad hay de formular las dos primeras? Antes de comenzar el interrogatorio, los dos jugadores saben que las dos primeras respuestas serán, forzosamente, "no". Entonces, ¿de qué forma pueden aportar información dichas preguntas? En apariencia, tras realizarse las dos primeras preguntas, los jugadores no han averiguado nada que no supieran ya; por consiguiente, parece que no se hallan más cerca de poder deducir su número de lo que ya estaban, y el juego no terminaría jamás. ¿En qué forma podrá la declamación ritual de noes, que ambos jugadores saben que forzosamente ha de ocurrir, acortar el número de preguntas necesarias para llegar a una respuesta afirmativa? El razonamiento parece impecable.

Supongamos que se limita el conjunto de números a los naturales desde 1 hasta 100. Se escribe en una ficha cada par de números consecutivos (1,2; 2,3; ...; 99,100). El árbitro toma una ficha al azar, pone sus números en la frente de dos jugadores racionales, y propone el juego siguiente. El jugador de número más bajo, k, ha de pagar k pesetas a su oponente. El árbitro le pregunta al jugador A si desea jugar, y a continuación, hace la misma pregunta al jugador B. Se efectúa el pago tan sólo cuando ambos jugadores dicen que sí.

Demostraremos ahora que el pago nunca llega a efectuarse. Si A ve el número 100 en su oponente, sabrá que su número es 99, por lo que contesta que no. Si ve el 99, razona así: "Yo tengo 98 o 100. Si tengo el 100, entonces B (por ser racional) dirá que no, y no habrá par-

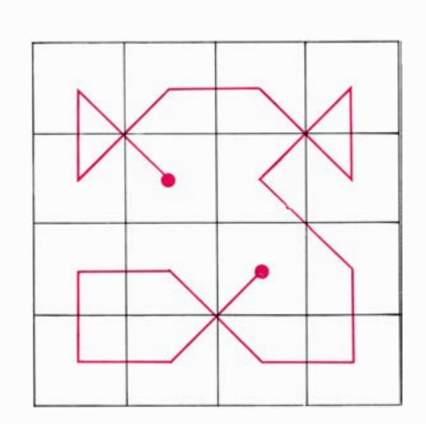
tida. Y si tengo el 98, evidentemente, no debo jugar. Por consiguiente, tengo que contestar "no". Si A ve el 98 razona así: "Tengo el 97 o el 99. Si tengo el 99, entonces B no querrá jugar, por las razones anteriores. Y si tengo el 97, perderé. Así pues, diré que no". Y así en todos los casos inferiores, incluso si ve un 1. En efecto, si A ve un 1, sabe que ganaría, pero sabe también que si él dice que sí, B dirá que no.

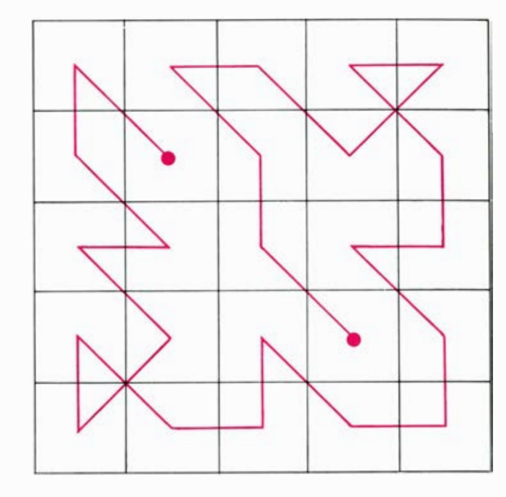
Supongamos que el conjunto de fichas sea infinito, sin cota superior para sus números. Demostraremos ahora que ambos hombres dirán que sí. A razonaría del modo siguiente: "Yo veo el número k. Mi número es, o bien k-1, o bien $k+1$. Si pierdo, perderé $k-1$ pesetas. Si gano, ganaré k pesetas. Ganar y perder son equiprobables, y como puedo ganar más de lo que arriesgo perder, el juego está a mi favor. Naturalmente, estoy de acuerdo en jugar". B razonará, por supuesto, de igual modo. Ahora bien, la situación es ridícula, porque el juego no puede ser favorable a los dos jugadores a la vez.

La paradoja queda formidablemente amplificada con una indicación de J. E. Littlewood, quien expone una versión de la paradoja en el primer capítulo de su *Mathematician's Miscellany.* Supongamos que haya 10 n copias de cada ficha, siendo n el menor de sus dos números. Así pues, 1,2 figura en 10 tarjetas, 2,3 en 100 tarjetas, 3,4 en 1000, etc. El juego se desarrolla como antes. Un jugador que vea en su oponente el número n, sabe que hay diez veces más fichas que tengan $n+1$ que $n-1$. Así, además de que ganar significa una peseta más que perder, la probabilidad de ganar parece ser ¡diez veces mayor para cada jugador! Littlewood atribuye esta monstruosa hipótesis al físico Erwin Schrödinger.

No resolveré esta paradoja el mes próximo, porque no estoy seguro de saber exactamente cómo hacerlo. No podré contestar a los lectores que me envíen opiniones, pero sí informaré más tarde acerca de sus comunicaciones. No es suficiente demostrar que el juego es justo exhibiendo su matriz de pagos. Evidentemente, el juego está equilibrado para ambas partes. El problema consiste en explicar por qué son defectuosos los razonamientos de A y de B.

John Horton Conway, de la Universidad de Cambridge, ha introducido en este juego una generalización profunda y desconcertante. Es fácil generalizarlo para n jugadores y n números consecutivos; lo que hace Conway es suprimir la condición de que los números hayan de ser consecutivos. Se permite ahora poner en la frente de n jugadores, todos honestos y racionales, cualquier entero no negativo (0 inclusive). En una pizarra, a la vista de todos los jugadores, se escriben m enteros no negativos diferentes, de los que uno es la suma de los números de los jugadores. Cada jugador puede ver los números de sus oponentes, pero no el suyo propio. El árbitro pregunta por turno a los jugadores: "¿Puede Ud. deducir el número que lleva?" El interrogatorio continúa cíclicamente hasta que un jugador concluye el juego diciendo que sí.



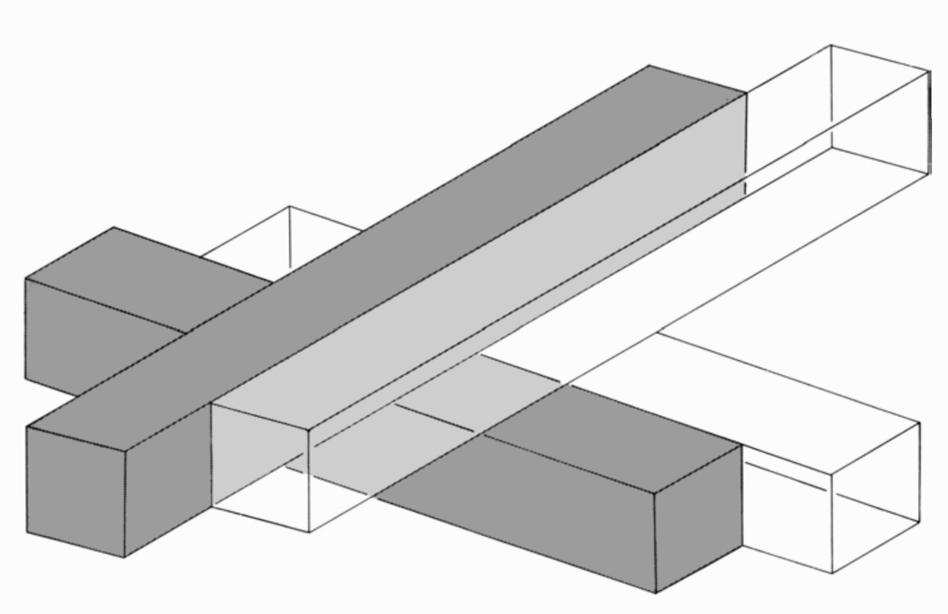


Soluciones de los recorridos de rey

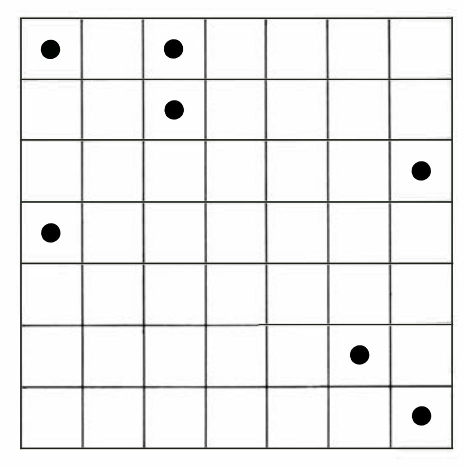
Conway ha demostrado el siguiente y notable teorema: Si m, número de sumas de la pizarra, no es mayor que n, el juego ha de tener final. Por ejemplo, supongamos que cada jugador tenga un 2, y que las sumas de la pizarra sean 6, 7 y 8. Con tres jugadores, Conway afirma que el juego terminará con una respuesta afirmativa a la 14-ava pregunta.

Establecer un algoritmo general que permita calcular cuándo terminarán estos juegos es, exceptuados ciertos conjuntos especiales de números, de una dificultad extraordinaria, y está lejos de hallarse resuelto. Escribe Conway: "Se entra en un círculo vicioso de la forma "A sabe que B sabe que C sabe que B sabe que C sabe...", incluso antes de formularse la primera pregunta, por lo que es muy difícil evaluar la cantidad de información disponible para cada jugador. De hecho, durante algún tiempo llegué a pensar que estas consideraciones podrían hacer que el juego no estuviese bien definido, y que se incurriría en paradojas. Ahora no lo creo así. Lo que sé bien es que resulta fatalmente sencillo cometer errores al calcular la cantidad de información disponible".

El juego de Conway presenta la misma paradoja que hemos discutido anteriormente. En el ejemplo dado, es fácil



Solución del problema del policubo



Solución del problema de la colocación

probar que aún antes de que comience el juego, cada jugador puede predecir que las tres primeras respuestas serán noes, por lo que parece que dichas preguntas podrían suprimirse, ya que después de la primera ronda, los jugadores no estarán mejor informados que antes de comenzarla. Sin embargo, si se elimina la primera vuelta, puede hacerse el mismo razonamiento respecto de la segunda, y el juego no tendría fin. Explicaré el próximo mes por qué estas paradojas no constituyen auténticas dificultades.

En razón de que con frecuencia la inducción matemática adopta la forma de "reducción al caso precedente", terminaré con una broma ya vieja. En vista de que un novato de primer curso era incapaz de decidirse entre Física y Matemáticas, se le sometió a una prueba de aptitud. La prueba constaba de dos partes. En la primera, al estudiante se le llevaba a una habitación donde había un fregadero, una cocinilla de gas de un solo fuego, y, en el suelo, una tetera vacía. El problema consistía en hervir agua. El estudiante superaba la primera parte si podía llenar la tetera en el grifo del fregadero, encender el quemador de la cocinilla y poner la tetera al fuego.

Para la segunda parte del ejercicio, el mismo estudiante era llevado a la misma habitación, pero ahora la tetera estaba llena de agua, sobre la cocinilla apagada. El problema seguía siendo hervir agua. Un físico en potencia se limitaría a encender el fuego. Un matemático en potencia vaciaría primero la tetera, y después, la pondría en el suelo. El problema se reducía así al caso anterior, que ya había sabido resolver.

Las soluciones de los problemas del mes pasado son las siguientes:

1. Salvo simetrías, la única solución del problema de las 15 bolas de billar americano es la mostrada en la ilustración de la página 105.

El coronel George Sicherman, de la State University of New York en Buffalo, que fue el inventor del problema, descubrió mediante computador que no existen soluciones para triángulos semejantes de órdenes 6, 7 y 8. Sicherman descubrió también una sencilla condición de paridad que demuestra la inexistencia de soluciones para todos los órdenes 2^n-2, cuando n sea mayor que 2.

He aquí la demostración correspondiente al caso de orden 6, que es el mínimo. Llamemos a, b, c, d, e, f a la fila superior del triángulo. Como módulo 2, la adición y la sustracción son idénticas, podemos expresar los restantes números, módulo 2, por adición. Así, la segunda fila es $a+b$, $b+c$, $c+d$, $d+e$, $e+f$. La siguiente fila comienza $a+2b+c$, $b+2c+d$ ... Continuando así, se llega al número del vértice inferior, que es $a+5b+10c+10d+5e+f$. El triángulo contiene 6 veces a, 20 b, 34 c, 34 d, 20 e y 6f. Todos estos números son pares, por lo que el triángulo es de paridad par. Sin embargo, el triángulo contiene 10 números pares y 11 impares, lo que le confiere paridad impar, y tenemos una contradicción.

La serie antedicha de números (6, 20, 34, 34, 20, 6) es la misma que la séptima fila del triángulo de Pascal (1, 7, 21, 35, 35, 21, 7, 1), disminuidos sus números en una unidad. La demostración general de Sicherman descansa en el conocido teorema que afirma que solamente las filas del triángulo de Pascal que ocupan lugar 2^n-1 están totalmente formadas por números impares.

Charles W. Trigg ha demostrado, entre otras cosas, que todo triángulo de diferencias absolutas de números consecutivos tendrá necesariamente un 1 en su vértice inferior. Además, ha expuesto la conjetura de que no existen otros triángulos aparte de los 11 expuestos este mes y el pasado. Quise gastar una broma a los lectores pidiéndoles que construyeran un triángulo de diferencias con 15 bolas marcadas con los números pares de 2 a 30. La única solución se deduce inmediatamente de la aquí expuesta, duplicando cada número.

2. Hay dos modos topológicamente distintos de que un toro se encuentre dentro de otro: el toro interior puede rodear el agujero del segundo, o no rodearlo.

Si los dos toros están enlazados, y uno de ellos tiene "boca", es imposible que éste pueda tragarse al otro sin que el toro interior rodee el orificio del exterior. Este resultado puede demostrarse dibujando sobre cada toro una línea cerrada, de modo que ambas líneas formen dos eslabones de una cadena. Por mucho que se deformen, es imposible desencadenarlos sin romperlos. Si un toro pudiera engullir a otro sin que éste rodease el agujero del primero, podría vomitar por la boca al toro devorado, y los toros quedarían desenlazados. Pero entonces las dos líneas quedarían también desenlazadas. Siendo imposible desencadenarlas, el canibalismo del tipo descrito resulta asimismo imposible.

Sin embargo, el toro caníbal sí puede tragarse al otro de manera que el toro tragado esté dentro en el primero de los sentidos explicados. La ilustración de la página 106 explica cómo se hace. En el proceso, el toro caníbal tiene que volverse del revés.

Un buen método para comprender lo que sucede es imaginar que el toro A se encoge, hasta convertirse en una banda de pintura que rodea al toro B. Se vuelve del revés el toro A, sacándolo por su propia boca. Haciéndolo así, la banda pintada pasa al interior, rodeando el agujero de B. Expandiendo la banda de pintura hasta obtener de nuevo un toro, se tiene la figura final de la serie.

3. La ilustración inferior de la página 107 muestra cómo pueden adosarse dos policubos (uno transparente, el otro sombreado) congruentes. Por extensión de los extremos, cualquier número finito de estas piezas puede adosarse de esta forma, de suerte que cada par "se toque", en el sentido de que compartan una porción de superficie común, y no queden agujeros interiores. Extendidos al infinito, un número infinito de policubos congruentes de orden infinito pueden "tocarse" mutuamente.

Si se elimina la condición de congruencia, y se exige en cambio convexidad, se sabe desde aproximadamente 1900 que un número infinito de sólidos convexos no congruentes pueden tocarse. Se ignora si un número infinito de sólidos convexos congruentes podrán tocarse mutuamente; pero Scott Kim ha demostrado recientemente (aunque no lo ha publicado) cómo puede obtenerse una tal disposición con un número finito arbitrariamente grande de dichos sólidos.

4. Los cuatro problemas lógicos de Robert Smullyan se resuelven como sigue:

(1) O A dice la verdad o miente. Supongamos que dice la verdad. Entonces B es bueno, y dice la verdad cuando afirma que A no es bueno. En este caso, A dice la verdad, pero no es bueno.

Supongamos que A miente. Entonces B no es bueno. Sin embargo, B dice la verdad cuando afirma que A no es bueno. Así que en este caso B está diciendo la verdad, pero no es bueno.

(2) O bien B dice la verdad, o bien miente. Supongamos que dice la verdad.

Entonces A es malo y tiene que mentir cuando dice que B es bueno. En este caso B dice la verdad, pero no es bueno.

Supongamos que B miente. Entonces, con seguridad, B no es bueno; así que A tiene que estar mintiendo cuando dice que B es bueno. Como B está mintiendo, A no es malo. En este caso, A está mintiendo, pero no es malo.

(3) B es o bueno o malo. Supongamos que es bueno. Entonces A y C deben ser del mismo tipo; tal como dice B, C está mintiendo cuando dice que B es malo; por consiguiente C es malo. Si C es malo, también tiene que serlo A.

Supongamos que B sea malo. Entonces A y C han de ser diferentes. C dice la verdad cuando afirma que B es malo, así que C debe ser bueno. Como A y C son diferentes, A tiene que ser malo. En cualquier caso, A es malo.

(4) La solución que Smullyan da a este problema es bastante larga, por lo que me contentaré con ofrecer un resumen. A y B son, o bueno-bueno, o malo-malo, o bueno-malo o malo-bueno. En cualquier caso, se ve en el análisis que tanto si C es bueno como si es malo, tiene que contestar "sí".

5. Se muestran soluciones de los dos recorridos de rey perdido en la ilustración superior de la página 107. El primer recorrido es único. Se desconoce si el segundo es único o si existe una solución con una intersección menos.

6. Aunque el problema de la elipse se presentó para el triángulo de lados 3, 4 y 5, lo resolveré para un triángulo cualquiera.

La elipse máxima que puede inscribirse en un triángulo equilátero es una circunferencia, y la elipse mínima que puede circunscribirse a un triángulo equilátero es también una circunferencia. Mediante proyección paralela podemos transformar un triángulo equilátero en un triángulo de forma arbitraria. Haciendolo así, las circunferencias inscrita y circunscrita se transforman en elipses no circulares.

La proyección paralela no altera la razón de las áreas del triángulo y de las dos líneas cerradas; así pues, las elipses resultantes tendrán las áreas máxima y mínima para cualquier triángulo producido por la proyección. Dicho de otro modo, la razón del área de la mínima elipse que pueda circunscribirse a un triángulo dado cualquiera al área de dicho triángulo es la misma que la razón del área de un círculo al área de un triángulo equilátero en él inscrito. Análogamente, la razón del área de la máxima elipse que puede inscribirse en un triángulo dado al área de dicho triángulo es la misma que la razón del área de un círculo al área de un triángulo equilátero circunscrito.

Es fácil demostrar que la razón del área del círculo inscrito al área del triángulo es $\pi/3\sqrt{3}$, y que la razón del círculo circunscrito al triángulo es 4 veces este número. Aplicando este resultado al triángulo de lados 3, 4 y 5, resulta que el área de la elipse máxima inscrita es de $2\pi/\sqrt{3}$, y que el área de la mínima elipse circunscrita es de $8\pi/\sqrt{3}$.

Los lectores que prefieran una demostración más formal, pueden hallarla en la obra de Heinrich Dörrie, *100 Great Problems in Elementary Mathematics* (Dover, 1965), páginas 178 y sigs.

7. La ilustración de la página anterior muestra el único modo (sin contar giros ni simetrías) de situar 7 fichas en una matriz de orden 7, de forma que todas las distancias entre pares de piezas sean diferentes.

8. La paradoja de la cuarta aleluya aparece al completarla mentalmente: "Cuyos ripios terminaban en el primer verso". Al completarla se contradice lo que el verso afirma. Las aleluyas del mes pasado indujeron al poeta cómico británico J. A. Lindon a improvisar otras, cuya versión libre damos aquí:

Un poeta moderno –el más infame–
ripios pergeñaba
que nadie soportaba.

"Al trabajo", por fin se dijo,
"ideas tengo hasta ser prolijo".
Tantas tuvo, que infame revoltijo
más que poema, aquéllo resultaba.
"Me voy yendo", dijo
Ideas tengo en la cocorota
Mas luego hallarlas me alborota".

Aún pudo ser peor
¡Dios no lo quiso!

Bastaba, simplemente,
leer primero
los versos más postreros.
Como realizaba
todo a la inversa,
había quien dudaba
si el poeta era hombre...
o viceversa.

Cada vez que intentaba recitarlos
–atrapar auditorio, ¡qué problema!–
todos preguntaban, "¿Qué pasa?".
"¿Es que no hay rima?".
Si sólo fuese rima...
¡Es que no hay tema!

Poeta el más veloz era Cuadrado:
Todos sus ripios salían pareados.

Distraído, al pobre Segismundo,
que errabundo, pulía el cuarto verso,
un camión le saco el reverso.
¡Desde entonces, el poeta Pintos,
escribe solamente versos quintos!

Taller y laboratorio

Planear suspendido o esquiar por el cielo a baja velocidad con un ala de excelentes características

C. L. Stong

En los últimos años, un número creciente de pilotos de planeador no profesionales se han unido a las gaviotas y a los pelícanos, en el juego de remontar y volar a pocos metros de altura sobre las inclinadas playas de las costas. Tierra adentro pasan rozando –en la dirección del viento– las dunas arenosas y planean por las laderas y colinas libres de obstáculos. Este deporte, en rápida expansión, conocido como "esquí aéreo o planeo suspendido" exige del entusiasta habilidades casi iguales como piloto y como entendido en aerodinámica.

Los aparatos que sirven de base a este deporte varían mucho en cuanto a detalles de diseño. Suelen costar muy poco, son casi siempre de fabricación casera, y se construyen con materiales ligeros. El piloto va colgado por unos aparejos.

Muchos planeadores reconocen su origen común en una cometa triangular, patentada en 1951 por Francis M. Rogallo y su mujer Gertrude. La cometa tiene un coeficiente de planeo de más de 4:1, lo que quiere decir que, en vuelo, avanza más de cuatro metros horizontalmente por cada uno que descienda. La mayoría de los pilotos aprenden primero en una cometa tipo Rogallo, a la que con frecuencia se la llama simplemente "ala". Muchos de ellos continúan volando con la cometa prefiriéndola a los planeadores colgantes que se crearon posteriormente. Estos últimos tienen unas características mejores que las cometas, pero son más caros y más peligrosos.

La cometa Rogallo fue concebida dentro de un costosísimo programa de investigación que buscaba la manera de sustituir el paracaídas a la hora de hacer descender a la tierra las cápsulas espaciales una vez que habían entrado en la atmósfera. Tal como fue desarrollado por los Rogallo, la cometa era un ala flexible en forma de delta. Se desplegaba y mantenía su forma gracias a un sistema de cables tensos, tal como un paracaídas mantiene su forma mediante sus cables de cobertura.

La National Aeronautics and Space Administration (NASA) abandonó la cometa como sistema para el aterrizaje de cápsulas espaciales debido, principalmente, a sus dificultades para mantenerlo empaquetado y desplegarlo después. Pero varias personas interesadas en la aerodinámica idearon rápidamente algunas modificaciones que convirtieron el artefacto en un útil "planeador suspendido". Uno de ellos fue M. A. Markovski, ingeniero aeronáutico que había participado en el diseño del CD-10, de la Douglas Aircraft Company, y que luego realizó investigaciones para la Sikorsky Aircraft. Markowski sustituyó algunas líneas tensoras de la cometa de Rogallo por un armazón de tubos de aluminio y equipó la estructura con una barra de control de aluminio y en forma de *Y* invertida. Un aparejo mantiene suspendido al piloto, ya sea sentado, ya sea boca abajo, posición desde la cual puede asir la barra de control y accionarla, desplazando su peso, para controlar el cabeceo, los giros y los guiños del aparato. Markowski describe así la evolución de su proyecto.

Mi interés por los planeadores que pueden hacerse despegar corriendo, comenzó como un pasatiempo, exclusivamente por diversión. Después de diseñar diversos biplanos de ala rígida y un monoplano de los del tipo de ala rápida, llegué a la conclusión de que el "ala" de Rogallo era el medio más práctico de adquirir experiencia en el arte del planeo suspendido. Siguiendo los consejos de varios informes técnicos, recopilados por la NASA, sobre el "ala" de Rogallo, diseñé un modelo a escala natural que por fortuna voló según estaba previsto. Sobrevolar rozando

Michael A. Markowski despegando en su EAGLE III

una pista de esquí cercana resultó ser una diversión fantástica y excitante.

Rápidamente dominé el arte de despegar y aterrizar sobre mis pies sin necesidad de instructor. Es una sensación increíble correr por el suelo unos cuantos pasos y conseguir elevarse sin esfuerzo por el aire. Apenas me daba cuenta de que estaba atado a la cometa, porque el aparato representa sólo el 20 por ciento del peso total del conjunto piloto-aparato. Suspendido boca abajo, el piloto tiene una sensación de absoluta libertad, quizá porque divisa el suelo a vista de pájaro.

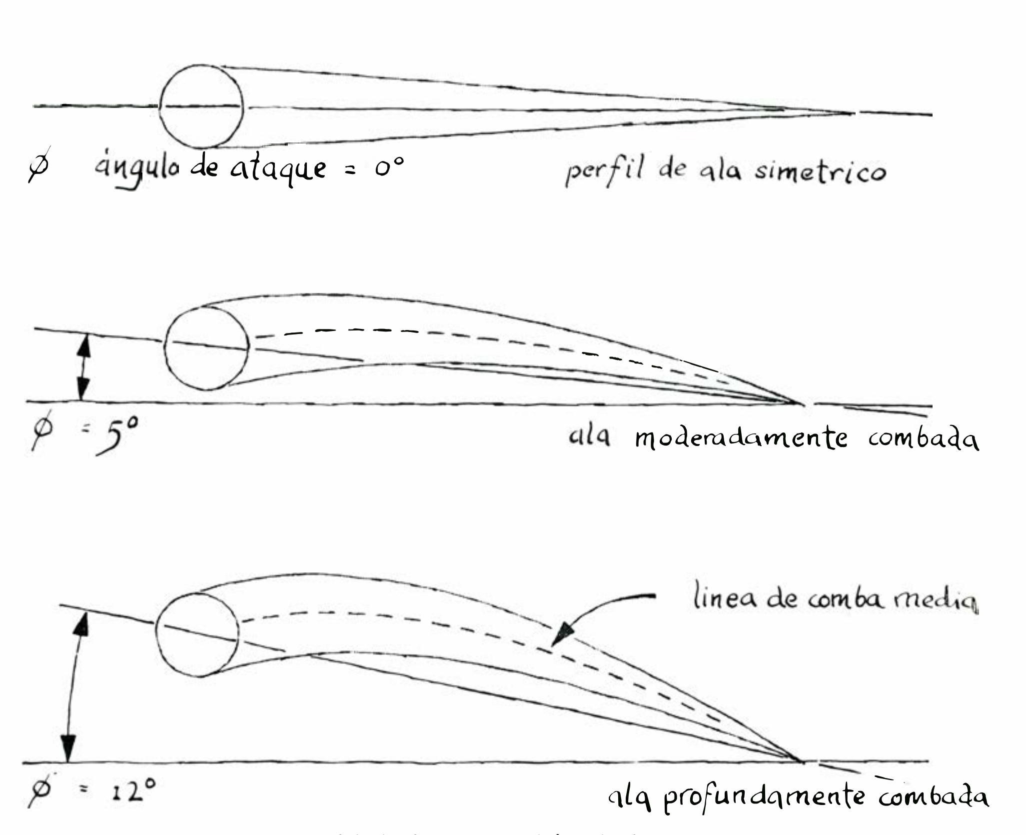
Mi primer planeador suspendido tenía una vela de polietileno de 0,1 mm y su peso total vacío era de 18 kilogramos. Me sirvió bien para muchos vuelos rasantes al terreno, hasta que la vela comenzó a estirarse, lo cual hizo disminuir el coeficiente de planeo. Conforme descendía la temperatura con la llegada del invierno de 1971, el plástico comenzó a romperse. Después de consultar a un montador de paracaídas, experto en telas, sustituí el polietileno por un nylon no desgarrable. Este material sirvió durante aproximadamente un año y después lo reemplacé, ya que el nylon resultaba demasiado elástico para mantener la curva de una buena vela. Ahora uso lona Dacron.

Este prototipo resultó ser el primero de una serie de modelos: le siguieron otros dos, uno conducido por control remoto y otro de vuelo libre, ambos en tamaño reducido; los utilicé asimismo para investigar a fondo otras características del "ala" de Rogallo. Por fortuna, estos experimentos me alertaron de la posibilidad de estrellarme. Cuando desciende el pico, aumenta la velocidad del aparato y la vela comienza a orzar y a flamear por el borde posterior. Las sacudidas de la vela avanzan hacia la cúspide y el aparato cae en picado en ángulo constantemente creciente hasta que choca contra el suelo.

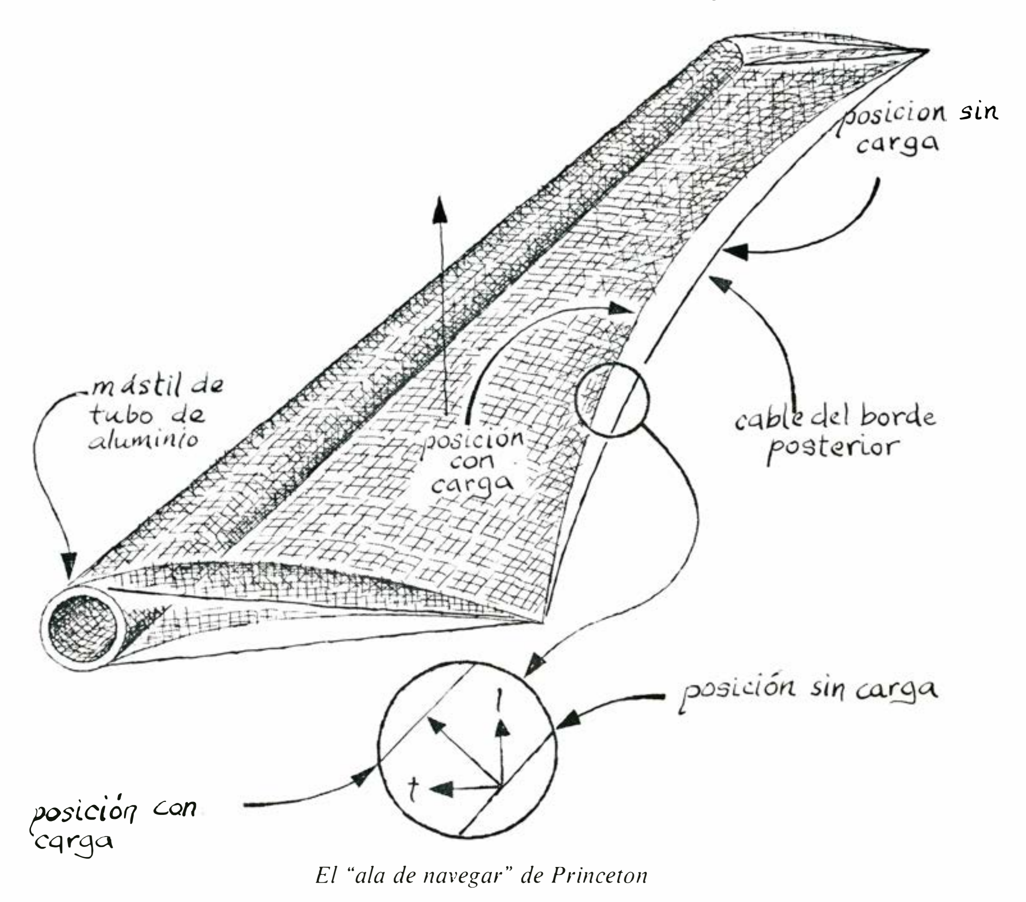
Habiendo observado este desconcertante comportamiento en los modelos pequeños, decidí comprobar personalmente si el fenómeno se presentaría también en planeadores de tamaño natural. Así fue. La primera solución que se me vino a la mente fue incorporar un estabilizador horizontal. Lo intenté con uno, pero aunque funcionó, su comportamiento era extraño y se rompió en seguida con el uso. La solución final consistió en suspender un peso a cierta distancia bajo la quilla del planeador. Este sistema funcionaba bien y no comportaba problemas estructurales.

El esquí aéreo es un arte y sólo puede dominarse únicamente tras una práctica continua. El piloto de estos planeadores ha de ser necesariamente autodidacta. El aparato no tiene un doble mando para el instructor, como en los aeroplanos y planeadores. Se comienza aprendiendo ciertos detalles complementarios sobre las propias facultades físicas, así como la forma de reacción de la cometa a la acción del viento al nivel del suelo, antes de aventurarse –aunque sea por poco tiempo– por los aires. Como los pajarillos, hay que aprender a usar las alas antes de abandonar el "nido".

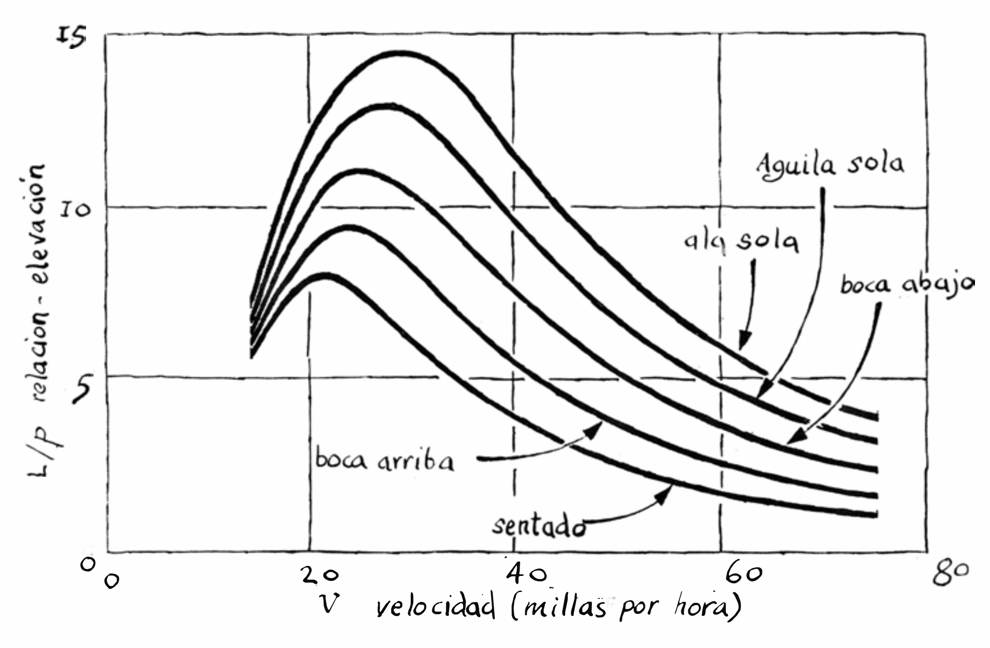
La cometa de Rogallo con la barra de



Respuesta del ala de navegar al ángulo de ataque



El "ala de navegar" de Princeton



Relaciones de elevación-resistencia en un EAGLE III

control y un aparejo seguro puede volar casi en cualquier sitio en que haya una pendiente adecuada que enfrente los vientos dominantes. Se encarece al principiante que comience su adiestramiento al pie de una pequeña colina cubierta de hierba o de una duna de arena con una inclinación de unos 25 grados. Es necesario que haya una brisa ascendente y continua de unos 15 kilómetros por hora. Conviene cerciorarse de que la brisa carece de ráfagas, ya que éstas pueden arrojarlo contra el suelo o bien elevarlo súbitamente 6 metros o más.

Cuando haya ensamblado la cometa cerca del pie de la pendiente, colóquela situando el pico en tierra y en dirección al viento. Dése una vuelta alrededor de la cometa comprobando una a una las partes metálicas para asegurarse de que están bien unidas y que no están dobladas, rotas o agrietadas. Al mismo tiempo compruebe la tela por si tiene agujeros, roturas o jirones. Asegúrese de que todos los puntos de amarre de la tela estén sanos y que los ojales metálicos no se hayan salido. Todos los aparejos deben estar lo suficientemente tirantes como para que restallen al estirarlos.

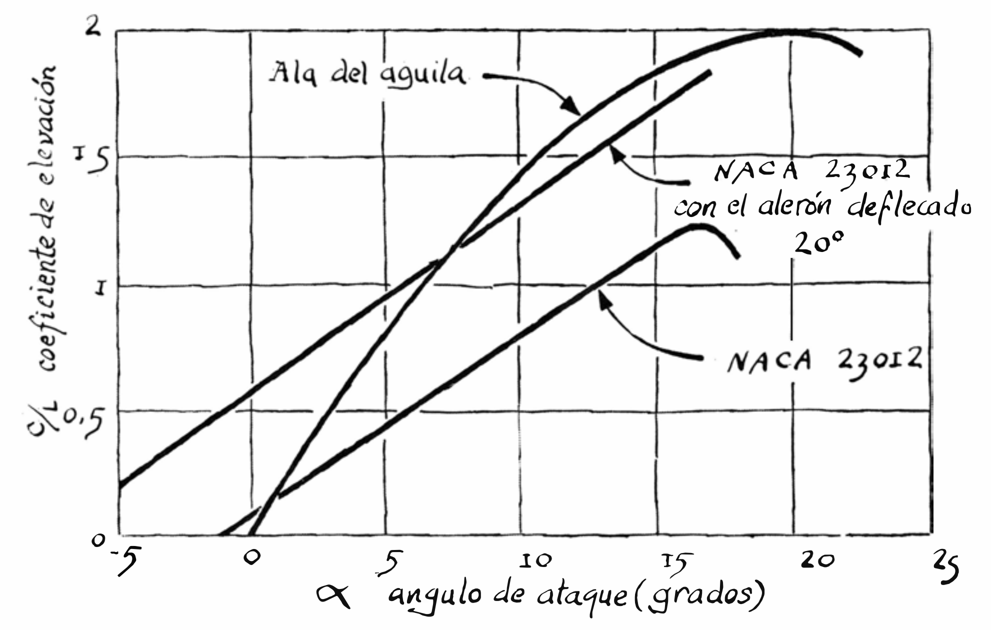
Cuando haya inspeccionado el planeador tome las barras verticales del sistema de control, levante la cometa por encima de su cabeza y corra apuntando contra el viento. Cuando vaya corriendo mueva el pico de la cometa hacia arriba y hacia abajo formando ángulos cada vez mayores, para notar qué efecto se siente. Al levantar el pico, hará que la tela se infle, recoja más viento y tire hacia arriba, con lo cual le costará más trabajo correr. Al bajar el pico, ocurre el efecto contrario: se quiebra la resistencia del viento y se podrá así correr más de prisa. Tenga en cuenta, sin embargo, que si baja demasiado el pico hará que la vela orce, con lo cual el viento actuará sobre la parte alta de la misma y empujará el aparato hacia abajo.

Continúe practicando a ras del suelo hasta que pueda predecir con certeza y "sentir" exactamente cómo las "alas" reaccionan a los distintos ángulos a los que usted sostenga la cometa. Cuando haya conseguido lo anterior, átese con las correas al aparato y continúe practicando hasta que el aparejo se le haga familiar.

En este momento puede comenzar a subir por la pendiente. Al principio, sin embargo, llegue hasta una elevación de sólo un metro aproximadamente. *Mantenga siempre el pico de la cometa apuntando directamente contra el viento,* incluso cuando suba con la cometa cuesta arriba. Conforme adquiera destreza, el viento le ayudará a subir la pendiente. Cuando corra cuesta abajo mantenga el armazón de control cerca del fondo, de manera que los montantes pasen cerca de sus hombros.

Por último, desde una elevación de 3 o 4 metros, comience a correr hacia abajo con la vela ligeramente inflada. Según vaya adquiriendo velocidad, empuje el armazón de control apartándolo un poco, elevando así el pico de la cometa. Si ha alcanzado velocidad suficiente notará que usted y la cometa se elevan por los aires. De lo contrario, sólo se frenará su carrera; en cuyo caso continúe hasta que durante la misma logre la correcta combinación de velocidad y ángulo de ataque para volar. Pronto estará rasando el suelo.

Una vez que haya conseguido volar, mantenga el equilibrio entre la proa y la popa desplazando su peso. Empuje contra la barra de control para mover su cuerpo hacia atrás, lo cual aumentará el ángulo de ataque, disminuyendo

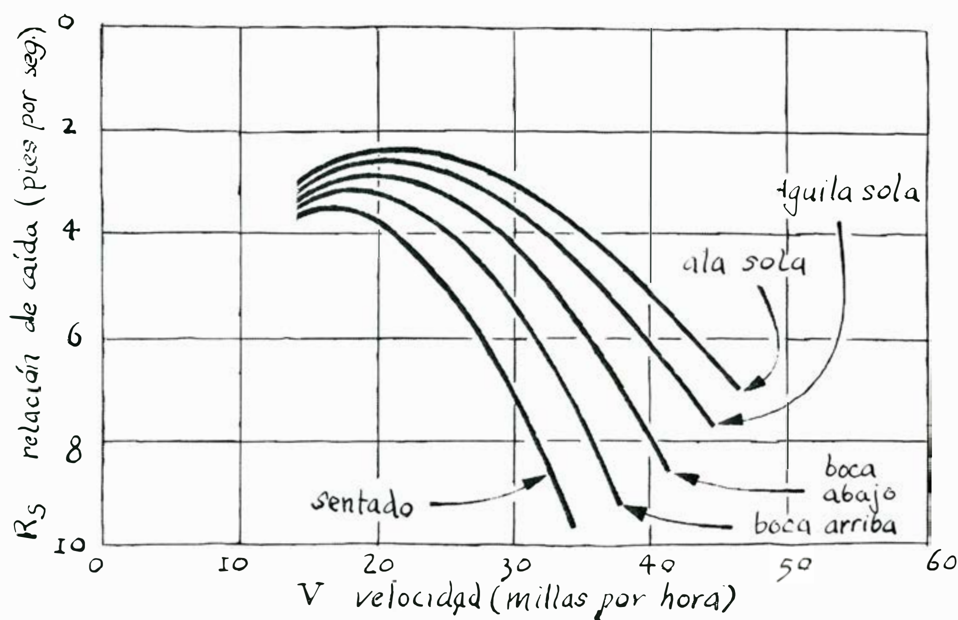


Características aerodinámicas de varias alas

así su velocidad. Tire de la barra de control para llevar su cuerpo hacia adelante, y disminuirá el ángulo de ataque e incrementará su velocidad. Mueva su cuerpo hacia la derecha ejerciendo una fuerza lateral contra la barra de control y hará que la cometa gire a la derecha; haga lo propio hacia el otro lado y obligará a que gire a la izquierda.

Cuando esté haciendo "esquí aéreo" a sólo un metro por encima de la superficie, recuerde que es mejor equivocarse en mantener el pico de la cometa alto y volar con excesiva lentitud que ponerlo hacia abajo, lo cual hará que el aparato caiga en picado exponiéndose a chocar bruscamente contra el suelo. Continúe practicando cerca de la base de la colina hasta que hacer vuelo rasante sobre el suelo le resulte tan natural como andar. Cuando haya conseguido esto, practique más y vaya subiendo gradualmente hasta la cumbre.

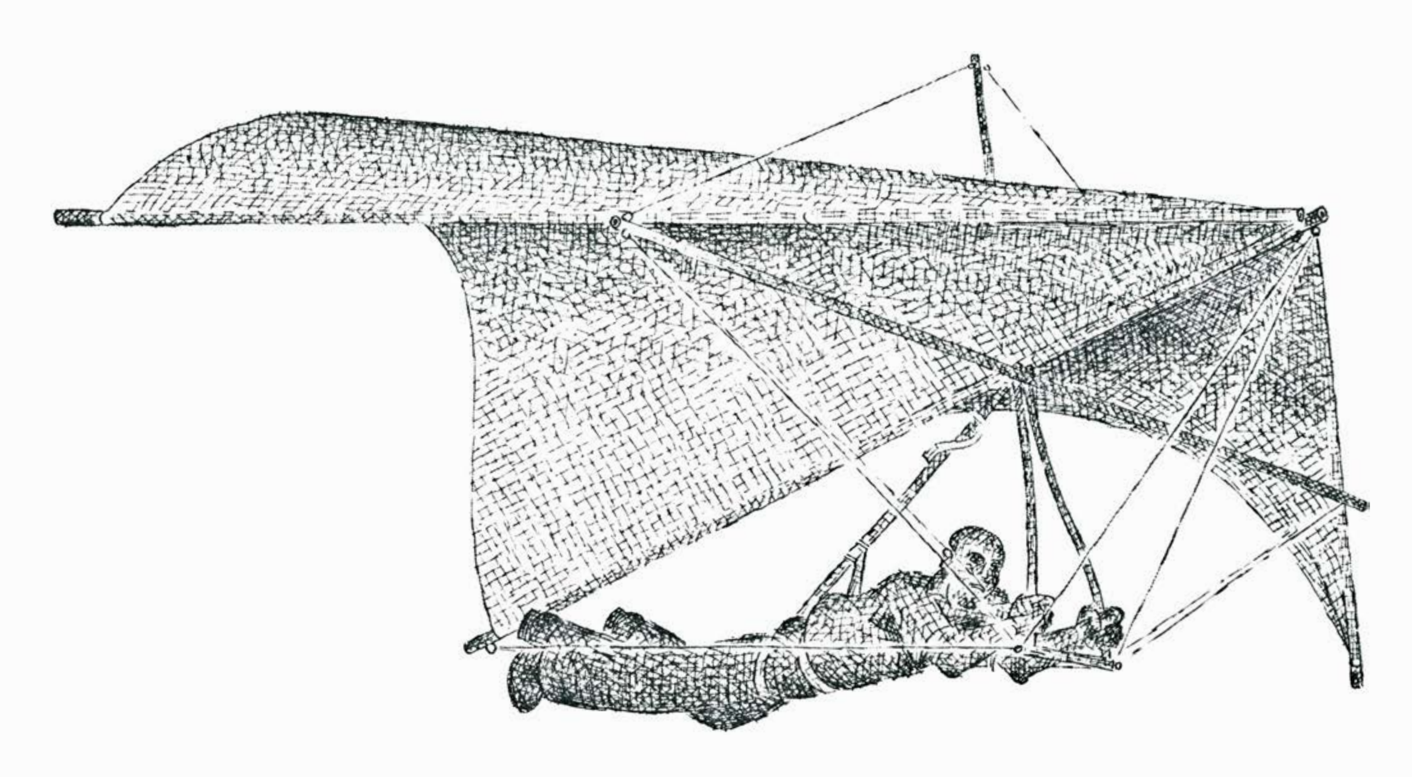
Concéntrese primero en aprender a mantener el equilibrio entre la proa y la popa, esto es, a cabecear. Si su aparato comienza a desviarse oblicuamente hacia un lado, descienda despacio y aterrice en cuanto pueda, para evitar ser arrastrado por la pendiente. Cuando haya conseguido dominar los movimientos de balanceo de proa y de popa, comience a practicar los giros. Sus primeros giros deberán ser tranquilos, suaves y de gran radio. Desplace su peso empujando muy suavemente hacia un lado las barras de control.



Indice de descenso vertical del EAGLE III

Los giros determinan que la cometa pierda altura más rápidamente que cuando se vuela en línea recta. Esto se puede compensar incrementando ligeramente el ángulo de ataque durante el giro. Evite levantar el pico de la cometa; no vaya a ser que coja un ángulo tal que haga que el planeador se detenga, pierda

Volando boca abajo



La cometa de Rogallo modificada

velocidad de vuelo y caiga al suelo. El ángulo de ataque óptimo sólo se descubre tras muchos intentos.

Conforme haya progresado hacia mayores altitudes tendrá ocasión de hacer piruetas y de maniobrar en las tres dimensiones del espacio, pero no precipite su aprendizaje. De un modo u otro se ha venido diciendo que la cometa de Rogallo es un aparato excepcionalmente seguro; de hecho, su seguridad es sólo relativa. Aerodinámicamente, el ala se caracteriza por la suavidad de su entrada en pérdida; es decir, cuando la cometa comienza a perder velocidad en vuelo, tiende a asentarse rápidamente, bajando la cabeza despacio en lugar de descender en picado. Su comportamiento se rige por las mismas leyes físicas que afectan a las demás máquinas voladoras.

Con la cometa Rogallo resulta más fácil volar que con otros aparatos, ya que el piloto controla el vuelo balanceando su propio cuerpo, acción que es más o menos instintiva. Sin embargo, el balanceo del cuerpo como técnica de control tiene sus limitaciones; ya que las fuerzas que se emplean para controlar son constantes mientras que las fuerzas perturbadoras varían con el cuadrado de la velocidad. Los controles aerodinámicos semejantes a los creados por los hermanos Wright permiten volar seguro en unas condiciones de viento en las cuales sería peligroso hacerlo con un planeador suspendido del tipo Rogallo. En otras palabras, las alas tipo Rogallo no son ningún juguete y pueden resultar mortales. Muchos principiantes no saben cómo aterrizar con seguridad cuando entran en pérdida a una altura de, por ejemplo, 12 metros. Esta maniobra puede y debe aprenderse practicando con paciencia.

Tanto la carga alar como la velocidad del aire son factores importantes en el comportamiento de todos los aparatos voladores. Teniendo en cuenta estas limitaciones, diseñé una serie de planeadores, basados en el ala de Rogallo, para pilotos con un peso de 45 a 95 kilogramos. Los planeadores podían volar con vientos de velocidades comprendidas entre menos de once y más de 32 kilómetros por hora. En general, los aparatos tienen una relación máxima de planeo de alrededor de 4,5:1 y una velocidad mínima de descenso vertical de 137 metros por minuto.

Aunque se puede disfrutar mucho con un planeador de estas características, la duración de los vuelos está limitada necesariamente a minutos más que a horas. A este respecto, planear con un ala Rogallo es comparable a cabalgar sobre las olas en la playa con un patín de "surf". Se disfruta sólo con la serie de viajes de regreso.

Buscaba un planeador de mejores características que me permitiese volar continuamente a lo largo de pendientes del lado del viento, y en ambos sentidos, haciendo giros de 180 grados en los extremos. También quería disfrutar de estas mejoras con un mínimo de problemas, lo cual significaba elaborar una máquina que una persona pudiera no sólo montar y desmontar, sino también plegar y cargar sobre un automóvil para su transporte.

Me vino a la mente una invención que prometía hacer posible tal planeador, al recordar un curso universitario de aerodinámica de bajas velocidades. Durante el mismo se describió un "ala de navegar" que había sido desarrollada en la Universidad de Princeton en 1948. Dicho dispositivo se había diseñado como vela perfeccionada para barcos. En 1952 se la adaptó para su posible empleo como superficie auxiliar de elevación en vehículos que levitan por la reacción de chorros de aire arrojados contra el suelo.

El ala de navegar consiste básicamente en un mástil tubular que soporta el borde anterior de una envoltura de tela y un juego de botavaras cortas y rígidas en el pie y en la punta del mástil, entre los cuales se tensa un cable fino que forma el borde posterior del ala. La estructura puede plegarse y guardarse.

La aerodinámica del ala de navegar es, a la vez, simple e impresionante. Su comportamiento se aproxima al de las clásicas alas rígidas en lo que respecta a sustentación y resistencia aerodinámica. A un ángulo de ataque nulo, cuando el plano del ala de navegar coincide con el plano de su movimiento, el ala de navegar adopta una sección transversal simétrica que no proporciona elevación. Al aumentar el ángulo de ataque, las superficies del ala asumen una forma combada que proporciona un empuje ascendente. Además, esta comba se hace más pronunciada al aumentar el ángulo de ataque, efecto aerodinámico equivalente al producido por un alerón automático de ala. Un gráfico muestra que la fuerza ascensional resultante del ala tiene una pendiente más pronunciada hacia arriba que la de un ala "rígida" clásica.

El principal factor que determina el coeficiente de planeo de un avión es la relación entre la longitud del ala y su anchura. Esta relación se denomina coeficiente aparente. Sería incorrecto afirmar que al doblar el coeficiente aparente de un ala se dobla también el coeficiente de planeo de un avión, pero al aumentar el coeficiente aparente se mejora en gran medida el coeficiente de planeo. (Dicho sea de paso, el coeficiente de planeo es numéricamente igual a la relación entre elevación y resistencia aerodinámica. El coeficiente de resistencia aerodinámica inducida es igual a $C_D = C^2_L / \pi AR$, en donde C_D es el coeficiente de resistencia aerodinámica, C_L es el coeficiente de elevación y AR es el coeficiente aparente, que es igual a la longitud del ala dividida por su cuerda media o anchura.)

Después de revisar las tablas de características aerodinámicas de varios cientos de alas, llegué a la conclusión de que con un ala de baja velocidad el coeficiente de elevación debería aumentar a razón de aproximadamente el 7,5 por ciento del ángulo de ataque; que la pérdida de sustentación debería empezar a ser notoria a un coeficiente de elevación de 1,6; que cuanto mayor sea el diámetro del borde anterior más suave será la pérdida de sustentación y mayor el coeficiente de elevación; cuanto más profunda sea la comba, mayor será el coeficiente de elevación, y que el ápice de la comba debería distar del plano más de un tercio de la distancia entre el borde anterior del ala y el posterior.

Observé, con cierta sorpresa, que estas características describían exactamente los resultados de pruebas experimentales realizadas en el ala de navegar, según informes de Princeton y la NASA. Estaba claro que en el ala de navegar se daban buenas características en una estructura simple y plegable que debería funcionar como la superficie de elevación del planeador definitivo.

Rápidamente construí e hice volar una serie de modelos a escala, algunos de los cuales eran controlados por radio. El primer prototipo, al que llamé "EAGLE I", tenía una envergadura de 12 metros y pesaba 32 kilogramos. El ala descansaba en un plano simple, esto es, no formaba un ángulo diedro curvado hacia arriba. Su coeficiente aparente era 8:1, su área medía 18,58 metros cuadrados y no tenía inclinación hacia atrás. La cola estaba dotada con un timón clásico, un timón de altura y un estabilizador horizontal. Para el control lateral sustituí los alerones por deflectores aerodinámicos, que podían levantarse mediante cables de control para crear una resistencia aerodinámica en las proximidades de cualquiera de los dos extremos del ala.

Las primeras pruebas de tierra y de "inflamiento de vela" se realizaron siguiendo el mismo procedimiento que se aplica para aprender a volar con un ala de Rogallo. Cuando una ligera brisa elevaba la máquina, el ala de navegar tomaba exactamente el perfil esperado. Las pruebas de vuelo comenzaron en dunas bajas. Los primeros vuelos rasantes en tierra indicaron que el timón de altura proporcionaba un perfecto control del cabeceo; también me sirvieron para desplazar el arnés de suspensión hasta su punto de equilibrio correcto. El timón resultó ser de cierta utilidad para controlar el derrape, pero los deflectores aerodinámicos eran inútiles. La tela de las superficies de control se distorsionaba mucho bajo la carga, lo cual hacía necesario un nuevo diseño.

Por lo demás, EAGLE I era totalmente estable en vuelo y tenía un coeficiente de planeo de cerca de 10:1. Aun cuando aumenté el tamaño de los deflectores en un 50 por ciento, la modificación nunca fue sometida a prueba. Una inspección crítica del aparato indicó que se podían introducir numerosas simplificaciones de diseño en la construcción de su estructura. Decidí conseguir un control lateral por el sistema usado por los hermanos Wright, a saber, el alabeo del ala. El resultado fue un aparato de diseño mucho más limpio, que pesaba 28,6 kilogramos.

Las pruebas de vuelo demostraron que el aparato era un planeador de excelentes características. Di una zancada mientras me enfrentaba a una brisa de 16 kilómetros por hora, tiré de la barra de control y me elevé casi verticalmente. La sensación fue fantástica. Sin embargo, el sistema de control lateral seguía siendo ineficaz. Resultaba obvio que la envergadura y el área del ala eran excesivamente grandes para manejarlas en cualquier situación que no fuera de calma absoluta.

Las pruebas y modificaciones posteriores condujeron a la construcción del EAGLE II. Esta máquina pesaba 34 kilogramos y tenía una envergadura de 10,36 metros; el ala tenía una superficie de 14,68 metros cuadrados. La supresión de unos cuantos pernos hizo posible el que se pudiera plegar fácilmente el avión para su transporte en el techo de un coche. El EAGLE II tenía un juego de poleas, cables y una barra de control que permitía al típico piloto de cometa adaptarse fácilmente a este aparato de tan altas características. El ala, con un coeficiente aparente de 7,25:1, carecía de inclinación hacia atrás. En la prueba desarrolló un coeficiente de elevación superior a 2, con una suave pérdida de sustentación. La pérdida de sustentación podría describirse más correctamente como "blanda", esto es, un descenso al estilo de un paracaídas. Cuando se accionaba la barra de control hacia atrás, el ángulo de ataque y la comba del ala aumentaban simultáneamente. El resultado de esa acción era un descenso pronunciado aunque lento. La respuesta a los controles seguía siendo buena hasta entrar en pérdida, pero era lenta para la guiñada o derrape en la condición "blanda" de pérdida.

El EAGLE II tenía un estrechamiento gradual del ala de 3:1 y un ángulo diedro de 8 grados. Al principio equipé el aparato para un ángulo diedro de sólo 2 grados, pero tendía a deslizarse demasiado en el derrape. También fijé en 2 grados el alabeo negativo en las puntas de las alas para evitar la pérdida de sustentación en los extremos (los extremos de las alas están torcidos para reducir el ángulo en 2 grados). La resistencia de la estructura se comprobó con una carga equivalente a 6 veces la fuerza de la gravedad, con un factor de seguridad de 1,5.

La estructura de la cola consistía en un bastidor de tubo de aluminio que soportaba la tela de nylon irrompible. Con el tiempo se hizo evidente que este material se estiraba demasiado; con la comba exagerada resultante, disminuía

el coeficiente de planeo. Como ya he dicho, ahora utilizo una tela de Dacron estabilizado.

Tanto el timón clásico como el timón de altura estaban equilibrados aerodinámicamente, esto es, estaban abisagrados un poco por delante de un cuarto de la cuerda de la superficie de mando. El estabilizador horizontal y el timón de altura podían desmontarse como una sola pieza de la quilla de la armadura. Las fuerzas desarrolladas por el timón y por la estructura en ángulo diedro del ala se combinaban para producir un par de derrape-balance que ayudaba a girar fácilmente. Las estructuras articuladas de tubo de aluminio cubierto que formaban las puntas de las alas alabeaban ambas las alas y servían como alerones. Las llamé "alaberones". Acopladas directamente al timón, deflectaban diferencialmente los extremos de las alas justo lo necesario para producir giros coordinados.

Ya se han realizado pruebas de vuelo exhaustivas del EAGLE III. En teoría, el máximo coeficiente de planeo de la máquina sin piloto es bastante más que 10:1. Naturalmente que esta característica se pierde en parte con la presencia del piloto, que crea fuerzas de resistencia aerodinámica pero no de elevación. La cantidad de resistencia que introduce el piloto depende de su posición. Los cálculos indican que cuando el piloto vuela tendido boca abajo el coeficiente de planeo del EAGLE III se aproxima a 11:1. En la posición de sentado es casi de 8:1. Estos valores suponen una velocidad aerodinámica óptima de unos 43 kilómetros por hora. La velocidad de descenso vertical también varía según la magnitud de la resistencia creada por el piloto: es de unos 61 metros por minuto. Las características del EAGLE III son, por tanto, compatibles no sólo con un vuelo sostenido en playas con pendiente situada a barlovento y demás terrenos similares, sino también con un planeo por todo terreno.

El EAGLE III es esencialmente un monoplano ultraligero, de altas características con un ala de comba variable. El piloto puede hacer que el aparato entre en pérdida en vuelo, pero dicha pérdida es suave comparada con la de un ala normal como la NACA 23012. Los principiantes no deberían intentar volar un EAGLE III, pero los pilotos que hayan llegado a dominar el planeo suspendido con una cometa de Rogallo se adaptan muy fácilmente a este aparato de altas características.

Libros

Desde los primeros organismos hasta el arrasamiento nuclear

Philip Morrison y Juana M. Golpe

The Game of Disarmament: How the United States and Russia run the Arms Race, por Alva Myrdal. Pantheon Books. Alva Myrdal escribe "casi desesperada", con cierta ira y con un tanto de invencible esperanza. Aportando a la carrera de estadista una fama de erudición creativa y de gran categoría, la autora ha servido a su país, Suecia, como ministro de desarme, como diplomático en el extranjero y como líder tenaz y competente de los intríngulis de las conferencias y comités de las Naciones Unidas a lo largo de 16 años. Su veredicto es claro y frío: "Quienes tienen el poder no tienen voluntad de desarme". Las superpotencias acumulan el rendimiento y las opciones, y los demás países comparten la responsabilidad en sus órbitas menores. En la misma preocupación se da por sentado "como si todos hubiésemos sido condicionados para continuar viviendo con los peligros con que se nos amenaza... tal como vivimos con la inevitable catástrofe personal de nuestra propia muerte... La diferencia estriba en que... el suicidio nuclear colectivo puede evitarse".

El libro es, en gran medida, una especie de crónica diplomática analítica que expone "una historia de oportunidades perdidas", desde el punto álgido del Tratado de Prohibición Parcial de Pruebas (atómicas) de 1963 hasta las menguantes esperanzas de una docena de otros tratados, conferencias y propuestas. La carrera de armamentos continúa; casi como un fin en sí misma. Cada una de las dos superpotencias parecen estar motivadas por la esperanza simple, irrazonable e inalcanzable de que "no debe ceder el primer puesto ante nadie". La prohibición parcial de pruebas "difícilmente puede considerarse como una medida de desarme". Su utilidad primaria se refería a la salud pública; su efecto, la ansiedad mundial por la lluvia radiactiva, se apaciguó con excesiva facilidad ante el nuevo campo subterráneo de pruebas, lo que permitió el libre desarrollo cualitativo de las armas nucleares. Se trata de una "solución muy propia del avestruz".

Los estadistas europeos ven claramente qué puede esperarse de la competición táctica nuclear de las superpotencias, que tan adelantada está en Europa como teatro de operaciones: un escenario en el cual *sus* países se convierten en laboratorios de explosión de sus bombas nucleares, mientras los territorios nacionales de las superpotencias siguen siendo refugios sagrados contra la nube en forma de seta. Y sin embargo eligen "comprar estabilidad política actual imponiendo... espantosos riesgos... sobre sus vidas y su futuro" Los reactores nucleares constituyen hoy el blanco de ciertos movimientos políticos en los principales países de Europa occidental, pero las armas nucleares tácticas, cuya amenaza es probadamente superior en varios órdenes de magnitud, quedan fuera del debate público. "Definitivamente, no hay interés en una discusión pública" acerca de sus implicaciones o de la devastación que conllevan.

La historia se va repitiendo día tras día. Para la señora Myrdal esto es simplemente lo irracional con interés compuesto. No cabe la menor duda de que, en último análisis, tiene razón, pero tres décadas de experiencia exigen una explicación más consistente. Comienza a evidenciarse que los líderes no ven la necesidad de perder las ventajas políticas que la carrera ofrece a todo nivel, desde las rivalidades internas en los servicios públicos hasta el sentimiento, no declarado, de dominio. En su informe final de principios de este año, Donald Rumsfeld, secretario de Defensa del Presidente Ford llegó a escribir: "Por mucho que uno pudiese desear otra cosa, el punto de vista del público e incluso de ciertos sectores del gobierno tiende a verse influido no tanto por análisis detallados cuanto por índices estáticos de las fuerzas nucleares relativas, tales como mecanismos de lanzamiento, cabezas nucleares, megatonelaje, precisión, capacidad propulsora y otros aspectos semejantes". Allí se describe el juego; por otro lado puede que una voz susurre que quizá el sistema direccional más reciente y un conjunto de astutas estrategias podrían hacer posible que las armas nucleares rindan un menudeo, hagan ganar puntos de menor cuantía, sin llegar a la disuasión ni a la guerra global. Esto no es una burda irracionalidad, sino lo que hemos conocido desde hace largo tiempo como el realismo chalado (y desde hace aún más tiempo como "hybris"). La señora Myrdal da a entender claramente que estamos ahora en un período de transición; esa opinión se está asentando especialmente en los Estados Unidos, indudable líder de la carrera.

Una breve y atrayente introducción personal –a la mayoría de los lectores les encantaría que fuese más larga– nos respalda el tono cortante del título. La autora da testimonio de tres sucesos en los que las superpotencias amagaron el progreso. En el primero, después de la "conversión" de los rivales tras la crisis de los misiles para Cuba, los delegados no alineados de las conversaciones de Ginebra elaboraron un "esquema harto ingenioso" para salvar la brecha respecto del número de inspecciones *in situ*, que había impedido por largo tiempo un tratado de prohibición de las explosiones experimentales. Las dos superpotencias emplearon la presión diplomática en las capitales de los delegados, dando a entender que la iniciativa de los delegados estaba poniendo en peligro importantes negociaciones directas. El resultado fue la prohibición parcial de las pruebas.

Nuevamente, una vez más, en un "lujoso almuerzo", el delegado de los Estados Unidos culpó a la señora Myrdal de que los EE. UU. no pudiesen ratificar el Protocolo de Ginebra, de 1925, que prohíbe la guerra química y biológica. "La presión era enorme"; los Estados Unidos estaban usando herbicidas en Vietnam. La señora Myrdal se mantuvo en su posición, insistiendo en que la no ratificación era preferible a una ratificación con reservas por parte de los Estados Unidos. Afortunadamente, su propuesta ganó. El tercer ejemplo mostró un "negativismo inexorable" por parte de

la URSS en cuanto a la comprobación de armas biológicas que deban estar prohibidas.

Las armas nucleares no son, por cierto, el único problema. Cañones, misiles, aviones, barcos y tanques son la industria de desarrollo económico en el mundo. Por ejemplo, los aviones militares supersónicos vuelan hoy con la bandera de 39 países subdesarrollados del mundo. No podemos pasar por alto el aumento de crueldad en los armamentos, desde el bombardeo zonal a los gases, toxinas, napalm y las balas ligeras de gran velocidad que se fragmentan, giran y generan ondas de choque en la carne blanco. Las luchas son numerosas y escasas las victorias.

Y, no obstante, queda esperanza. Quizá la pasividad llegue a su fin. Podría surgir una conferencia mundial del desarme si cada una de las superpotencias le diese vida mediante dos compromisos: no ser la primera en usar armas nucleares, ni atacar con ellas a países que carezcan de las mismas. Los chinos han fijado estas condiciones como requisitos previos para participar en tal conferencia; a la señora Myrdal le parecen "columnas de confianza" para la puerta de entrada de cualquier conferencia. Tal conferencia tendría muchas tareas ante sí; ella cita ocho, desde los pasos dados hacia una disuasión mínima hasta la erradicación de bases en el extranjero. Los propios estados más débiles podrían fijar sus convenios.

Los estadounidenses deberían actuar en último término. Habría que ceder un lugar de prestigio en tales negociaciones a los expertos, porque la tecnología de ese país ha sido la principal maestranza armamentista.

La señora Myrdal sigue escribiendo que el pueblo ruso, por su dura experiencia, aborrece "indudablemente" más la guerra que el pueblo norteamericano. Pero en los Estados Unidos la gente puede escuchar, manifestarse. Afirma: "no espero que el libro sea traducido al ruso". (Pero también allí habrá quienes sopesarán sus argumentos). Y acerca de Estados Unidos: "No conozco ningún otro país más capaz de cambiar su línea política". Quizá esté llegando la época, como tiene que llegar para la supervivencia humana, en que nuestros líderes encuentren el valor y la confianza pública suficientes como para empezar en serio, de verdad, el largo camino que nos aleje del abismo abierto a nuestros pies. El peligro actual es bastante claro, y no se suscita muy lejos de las riberas del río Potomec (P. M.)

PALEONTOLOGÍA. T. 1: Parte General e Invertebrados (2.ª ed. revisada y ampliada). B. Meléndez. Paraninfo S. A. Madrid, 1977. Los enseñantes de paleontología general nos encontramos con una dificultad común: demasiada materia y poco tiempo. Se necesita, pues, una síntesis de los temas que permita dar una idea general de lo más interesante. Deben exponerse las ideas generales del reino vegetal y del reino animal y específicas –en cuanto a morfología, significación y modo de estudio se refiere– de los representantes fósiles de los mismos.

Se ha de hacer biología, en el sentido más amplio de la palabra. Unas veces, la neobiología nos ayuda enormemente, pero muchas obras hemos de ceñirnos a la paleobiología e intentar interpretar lo pasado y lo actual a partir de la misma, sin muchos nexos conocidos en no pocos casos. De ahí la oportunidad de un libro de paleontología sintetizado, reducido, que sirva para informar a aquellas personas interesadas en estos temas –y no sólo para los alumnos que los han de estudiar en los diversos centros–, que les ofrezca una visión de conjunto de esta ciencia con datos precisos y contrastados por la bibliografía actual correspondiente y que les pueda sugerir nuevos estudios, de tipo general o específicos.

Ya conocíamos otros libros del profesor Meléndez sobre esta materia, pero la aparición del que comentamos nos alegró enormemente, pues cumple los requisitos antes mencionados, expuestos desde un punto de vista pedagógico, claro y lúcido y expresados en unas setecientas páginas. La necesidad que cubre esta publicación es la falta de un libro de este tipo en lengua española. La bibliografía específica en nuestro idioma es abundante, pero no hay un tratado general asequible a todos los lectores. Se ha de recurrir a los textos extranjeros, lo que requiere el conocimiento de otras lenguas, y el alumnado universitario suele ignorar incluso lo que, entendemos, resulta imprescindible: una lengua latina del grupo no-español y otra sajona (sería muy conveniente poseer amplios conocimientos de latín y griego).

La presentación general del libro del profesor Meléndez es correcta y de fácil manejo; en resumen, de aspecto tipográfico agradable. Características a resaltar en una obra compendiosa, con gran riqueza de contenido. Ofrece una diversidad de tipos de letra que, al mismo tiempo que señaliza aspectos diversos de lo tratado, hace la lectura más llana. Las notas al pie quizá estarían mejor dentro del texto, salvo para citas bibliográficas concretas y etimologías, éstas abundantes en la obra. El "corpus" gráfico es muy bueno, con amplia aportación de datos españoles, citas y fósiles, algo que siempre alabamos en las obras del autor.

El libro podría dividirse en tres bloques de materias. El primero sería el conjunto de capítulos dedicados a explicar cuestiones de la paleontología general y a definir los conceptos más universales de la misma. Son temas bien expuestos y con abundante material gráfico, del que una tercera parte son citas españolas, lo que juzgamos realmente importante. Se incluyen documentos históricos, clásicos y modernos, con un valioso apéndice de la historia de la misma en España. Hay un apartado dedicado a las observaciones relativas a la búsqueda de los fósiles y a la manera de tratarlos antes de su estudio. Dentro de estas generalidades, los capítulos más significativos son los que configuran la interpretación de los restos del pasado en su sentido más amplio –definición, morfología, sistemática y funcionalismo del pretérito biológico y su entorno– y los que dan a conocer los métodos paleontológicos teóricos, prácticos y comparativos, así como la aplicación de estos conocimientos al estudio e interpretación de la evolución biológica y su ayuda a la estratigrafía histórica en la seriación de los organismos en las capas terrestres.

El autor dedica a la micropaleontología todo un capítulo, relativamente corto con respecto a las otras dos partes, toda vez que constituye por sí misma una parte importante de nuestra Ciencia. El autor avanza que estudiará detenidamente la micropaleozoología de invertebrados.

De hecho, ya se dan unas nociones generales de micropaleobotánica, pero creemos que quizá fuera conveniente ampliarlas en aquellos grupos más problemáticos como las algas pirrofitas y crisofitas, en la linde de los reinos vegetal y animal. Las primeras aparecen con muy pocas formas francamente vegetales por ser un verdadero nexo con protozoos flagelados; ahora bien, por realizar la función fotosintética se consideran plantas (si bien entre las histricosféridas y ebriedinas se registran algunas formas animales). Dentro de las crisofitas, las crisomonadales, por estar ligadas a zooflagelados y rizópodos, las cocolitoforales y sus afines las discoastéridas postcretácicas, aunque se consideren un grupo sistemáticamente artificial al estar

representadas desde el Cámbrico hasta la actualidad, y otras formas aún poco conocidas e incompletamente interpretadas, incluidas en estos mismos estadios problemáticos (silicoflageladas, ofioboláceas). En cuanto a la micropaleozoología de vertebrados, una alusión más explícita a los diversos temas que se pueden tratar en ella, incluyendo breves referencias a los mamíferos mesozoicos, o a los estudios microscópicos en diversas estructuras dentarias u osteológicas, por ejemplo, ayudaría a completar la visión de la micropaleontología.

El tercer gran grupo de temas, reservado al estudio de los invertebrados fósiles, ocupa más de la mitad del texto y está tratado de un modo actualizado. En un primer apartado, se contempla el despliegue evolutivo de los no artrópodos; a parte de sus morfologías y distribuciones en los tiempos geológicos, se obtiene con el estudio de estos primeros capítulos un conocimiento de las relaciones entre estos animales y de los rasgos generales que tendrían otros anteriores a ellos, desconocidos, pero que pudieron ser su origen. Los invertebrados artrópodos, en su concepto más amplio, incluyen no sólo las formas típica y estrictamente fósiles como los trilobitomorfos, sino también los restos conocidos de aquellos seres que tienen aún hoy representantes, aportando una documentación amplia de los escasos fósiles de algunos de estos grupos, siguiendo la historia de los mismos hasta nuestros días. El último capítulo del libro, consagrado al estudio de los equinodermos y homalozoos, está expuesto, aunque sintéticamente, con gran cantidad de datos, que permiten obtener la necesaria visión de este vasto conjunto: primitivas adaptaciones (tan diversas, surgidas remotamente y que sugieren un posible origen polifilético de los mismos) y diversificación posterior, originaria a su vez de nuevos grupos. Examina las distintas filogenias que se pueden establecer dentro de los equinodermos así como la posible ancestralidad de otros *phyla,* representada por organizaciones muy relacionadas con ellos, si suponemos al subfilum homolazoos (= carpoideos) en el origen de los cordados. Esto último queda bien patentizado en el apéndice que cierra el texto, compendiando claramente el estudio de este subfilum y de su problemática con respecto a su probable entronque con los procordados y que sirve de puente introductor al anunciado segundo tomo de la obra, tan esperado, sobre los vertebrados en sentido amplio y la paleobotánica (J.M.G.)

Bibliografía

Los lectores interesados en una mayor profundización de los temas expuestos pueden consultar los trabajos siguientes:

EMBALSES SUBTERRANEOS PARA EL CONTROL DEL CICLO DEL AGUA

Icebergs as a Fresh-Water Source: An Appraisal. W. F. Weeks y W. J. Campbell en *Journal of Glaciology*, vol. 12, n.º 65, págs. 207-233; 1973.

Ground Water in Perspective. Raymond L. Nace en *Water Resources Bulletin*, vol. 9, n.º 1, págs. 18-24; febrero, 1973.

RAMAPITHECUS

Newly Recognized Mandible of *Ramapithecus*. David Pilbeam en *Nature*, vol. 222, n.º 5198, págs. 1093-1094; junio 14, 1969.

Ramapithecus en *Primate Evolution: An Introduction to Man's Place in Nature*. Elwyn L. Simons. The Macmillan Company, 1972.

The Primate and Other Fauna from Fort Ternan, Kenya. Peter Andrews y Alan Walker en *Human Origins: Louis Leakey and the East African Evidence*, dirigido por Glynn Ll. Isaac y Elizabeth R. McCown. W. A. Benjamin, Inc., 1976.

SEMICONDUCTORES AMORFOS

Electronic Processes in Non-Crystalline Materials. N. F. Mott y E. A. David, Oxford University Press, 1971.

On State of Amorphous Threshold Switches. Kurt E. Petersen y David Adler en *Journal of Applied Physics*, vol. 47, n.º 1, págs. 256-263; enero, 1976.

Localized States in the Gap of Amorphous Semiconductors. Stanford R. Ovshinsky en *Physical Review Letters*, vol. 36, n.º 24, págs. 1469-1472; junio 14, 1976.

Valence-Alternation Model for Localized Gap States in Lone-Pair Semiconductors. Marc Kastner, David Adler y H. Fritzsche en *Physical Review Letters*, vol. 37, n.º 22, págs. 1504-1507; noviembre 29, 1976.

INMUNOLOGIA DEL CANCER

The G_{IX} System: A Cell Surface Allo-Antigen Associated with Murine Leukemia Virus; Implications Regarding Chromosomal Integration of the Viral Genome. Elisabeth Stockert, Lloyd J. Old y Edward A. Boyse en *The Journal of Experimental Medicine*, vol. 133, n.º 6, págs. 1334-1355; junio 1, 1971.

Serum-Mediated Leukemia Cell Destruction in AKR Mice: Role of Complement in the Phenomenon. Robert L. Kassel, Lloyd J. Old, Elizabeth A. Carswell, Nancy, C. Fiore y William D. Hardy, Jr., en *The Journal of Experimental Medicine*, vol. 138, n.º 4, págs. 925-938; octubre 1, 1973.

Cell Surfaces and Malignancy. Dirigido por Peter T. Mora. DHEW Publicación n.º (NIH) 75-796, U.S. Government Printing Office.

Immunological Surveillance against Neoplasia. George Klein en *The Harvey Lectures*, Series 69, págs. 71-102; 1975.

An Endotoxin-Induced Serum Factor That Causes Necrosis of Tumors. E. A. Carswell, L. J. Old, R. L. Kassel, S. Green, N. Fiore y B. Williamson en *Proceedings of the National Academy of Sciences of the United States of America*, vol. 72, n.º 9, págs. 3666-3670; septiembre, 1975.

LA DESAPARICION DE LAS MANCHAS SOLARES

The Maunder Minimum. John A. Eddy en *Science*, vol. 192, n.º 4245, págs. 1189-1202; junio 18, 1976.

LA INVESTIGACION DEL HERBARIO

Drugs and Foods from Little-known Plants: Notes in Harvard University Herbaria. Siri von Reis Altschul. Harvard University Press, 1973.

Empirical Aztec Medicine. Bernard Ortiz de Montellano en *Science*, vol. 188, n.º 4185, págs. 215-220; abril 18, 1975.

BIOQUIMICA DE LA SUBNORMALIDAD

Preventability of Perinatal Injury. K. Adamsons y H. A. y Fox, ed. A. R. Liss, Inc. N. Y., 1975.

Perspectives in Brain Research. M. A. Corner y D. F. Swaab, ed. Elsevier, Amsterdam, 1976.

Genetics: Voyage of Discovery for Everyman. Ch. R. Scriver, ed. Pediat. Res. 10, 865, 1976.

Prevention of Handicap of Perinatal Origin. M. Wynn y A. Wynn, London, 1976.

Brain Damage and Perinatal Biochemistry. F. Mayor en *Reflections on Biochemistry*, ed. A. Kornberg *et al.*, Pergamon Press, Oxford, 1976.

Polimorfismos Bioquimicos. F. Mayor, II Reunión Nacional de Genética Humana, Barcelona, abril, 1977.

Desarrollo pre y posnatal del Cerebro. F. Mayor, Real Academia de Farmacia, Madrid, mayo, 1977.

LA PARADOJA DE STEIN EN ESTADISTICA

Empirical Bayes Methods Applied to Estimating Fire Alarm Probabilities. Grace M. Carter y John E. Rolph en *Journal of the American Statistical Association*, vol. 69, n.º 348, págs. 880-885; diciembre, 1974.

Biased versus Unbiased Estimation. Bradley Efron en *Advances in Mathematics*, vol. 16, n.º 3, págs. 259-277; junio, 1975.

Data Analysis Using Stein's Estimator and Its Generalizations. Bradley Efron y Carl Morris en *Journal of the American Statistical Association*, vol. 70, n.º 350, págs. 311-319; junio, 1975.

JUEGOS MATEMATICOS

The Method of Mathematical Induction. I. S. Sominskii, traducido del ruso por Hallina Moss. Blaisdell Publishing Co., 1961.

Mathematical Induction. Albert A. Blank en *Enrichment Mathematics for High School*. National Council of Teachers of Mathematics, 1963.

TALLER Y LABORATORIO

Hang Gliding: The Basic Handbook of Skysurfing. Dan Poynter. Publicado por el autor, Box 4232, Santa Barbara, Calif., 1974.